

UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA  
Departamento de Lenguajes y Sistemas Informáticos  
*Escuela Técnica Superior de Ingeniería Informática*



---

**HARNESSING FOLKSONOMIES FOR RESOURCE  
CLASSIFICATION**

---

**PhD THESIS**

**Arkaitz Zubiaga Mendialdua**  
MSc in Computer Science  
2011



UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA  
Departamento de Lenguajes y Sistemas Informáticos  
*Escuela Técnica Superior de Ingeniería Informática*



## **HARNESSING FOLKSONOMIES FOR RESOURCE CLASSIFICATION**

**Arkaitz Zubiaga Mendialdua**

MSc in Computer Science, Mondragon Unibertsitatea

Advisors:

**Víctor Fresno Fernández**

Assistant Professor in the Lenguajes y Sistemas Informáticos Department at  
Universidad Nacional de Educación a Distancia

**Raquel Martínez Unanue**

Associate Professor in the Lenguajes y Sistemas Informáticos Department at  
Universidad Nacional de Educación a Distancia



©2011 Arkaitz Zubiaga Mendiakdua

This work is licensed under the

Creative Commons Attribution-ShareAlike 3.0 License.

To view a copy of this license, visit

<http://creativecommons.org/licenses/by-sa/3.0/>

or send a letter to

Creative Commons,

543 Howard Street, 5th Floor,

San Francisco, California, 94105, USA.



*"A free culture has been our past, but it will only be our future if we change the path we  
are on right now."*  
—Lawrence Lessig

*"Free software is a matter of liberty, not price. To understand the concept, you should  
think of free as in free speech, not as in free beer."*  
—Richard Stallman

*"The only valid censorship of ideas is the right of people not to listen."*  
—Tommy Smothers





## Acknowledgments

First and foremost I would like to earnestly thank my advisors, Víctor and Raquel, who have been helping me throughout this research. Their hard work guiding me has enabled me to acquire a deep insight into both research and the field of social media mining. Their guidance has been of vital importance for pursuing first the Master Thesis, and the PhD Thesis afterward. Since the moment I decided to work on the novel field of social media, which was also new to them, they responded with a very willing and excellent attitude to assist with this project.

I would also like to thank my colleagues at the University, with whom I have discussed many research matters, that allowed me to grow as a researcher. Furthermore, all the leisure times we enjoyed together were of utmost importance to help me feel welcome and comfortable in the department. I would especially like to thank my academic brother Alberto, who has worked alongside me. With Alberto I have learned lots of new things and together worked on significant research that was presented at an international conference.

I am also very grateful to Markus and Christian, from the Graz University of Technology, with whom I had the pleasure of cooperating. They kindly hosted me in Graz, Austria, during my research stay. The stay was helpful in many aspects, such as getting to know other researchers who work in the same field as I, and sharing our thoughts and knowledge. It was especially useful to produce a sound joint research work which was presented at a consolidated and prestigious international conference.

I deeply appreciate and want to thank Jake for his kindness and effort in reviewing this dissertation. Jake provided a handful of tips and suggestions to improve the English in this document.

I also want to thank all the friends who have been with me all this time. I would like to thank not only those in my hometown Arrasate in the Basque Coun-

try, with whom I spent most of my childhood, but also my friends in Madrid, who have helped me settle in the city.

Last, but not least, I would sincerely like to thank my parents, José Ramón and Maribi, and my sister Igone. Without them, I would not be who I am today. Their help and patience throughout so many years has been fundamental in helping me reach this goal. And finally, my most sincere thanks to Nayeli, with whom I have shared most of my time during this work. She has been patient and understanding throughout the process of writing this dissertation and I will always wholeheartedly appreciate her.

## **Institutional Acknowledgements**

The research presented in this work has been in part funded by the Regional Government of Madrid under the Research Networks MAVIR (S-0505/TIC-0267) and MA2VICMR (S-2009/TIC-1542), the Regional Ministry of Education of the Community of Madrid, and the Spanish Ministry of Science and Innovation projects QEAVis-Catiex (TIN2007-67581-C02-01) and Holopedia (TIN2010-21128-C02-01).

## Abstract

In our daily lives, organizing resources into a set of categories is a common task. Organizing resources into categories makes searching through those resources easier by limiting the focus to a specific category. Limiting the focus significantly reduces the amount of information one must search. Categorization becomes more useful as the collection of resources increases, when managing resources becomes more and more difficult if they are not organized appropriately. Large collections like those made up by books, movies, and web pages, for instance, are usually cataloged in libraries, organized in databases and classified in directories, respectively. However, the usual largeness of these collections requires a vast endeavor and an outrageous expense to organize manually.

Recent research is moving towards developing automated classifiers that reduce the increasing costs and effort of the task. Most of the research in this field has focused on self-content, where the publisher is the only author, as a data source to discover the aboutness of the resource. Self-content presents the problem that it is not always representative enough, and sometimes it is difficult to access depending on the type of resource. Little work has been done analyzing the appropriateness of and exploring how to harness the annotations provided by users on social tagging systems as a data source. Users on these systems save resources as bookmarks in a social environment by attaching annotations in the form of tags. It has been shown that these tags facilitate retrieval of resources not only for the annotators themselves but also for the whole community. Likewise, these tags provide meaningful metadata that refers to the content of the resources.

In this thesis, we deal with the utilization of these user-provided tags in search of the most accurate classification of resources as compared to expert-driven categorizations. After performing a set of experiments to choose a suitable classifier

for this kind of task, we explore social annotations looking for a way to best use them. For this purpose, we have created three large-scale datasets including tagging data for resources from well-known social tagging systems: Delicious, LibraryThing, and GoodReads. Those resources are accompanied by categorization data from sound and consolidated expert-driven taxonomies. From these resources the appropriateness of social tags for predicting categories can be evaluated.

Specifically, we first study several ways of representing the massive number of social tags by amalgamating the contributions of large communities of users. We analyze their suitability for the classification task, upon both broader top level categories and narrower deep level categories. Then, we explore the nature, characteristics, and distributions of tags in folksonomies, in order to determine how the settings of each system affect the tagging behavior and the usefulness of tags for the classification task. We go deeper into tag distributions by analyzing the usefulness of weighting schemes based on inverse frequency values. Finally, using state-of-the-art user behavior detection processes, we identify users on social tagging systems who better fit the classification task.

To the best of our knowledge, this is the first research work performing actual classification experiments utilizing social tags. By exploring the characteristics and nature of these systems and the underlying folksonomies, this thesis sheds new light on the way of getting the most out of social tags for the sake of automated resource classification tasks. Therefore, we believe that the contributions in this work are of utmost interest for future researchers in the field, as well as for the scientific community in order to better understand these systems and further utilize the knowledge garnered from social tags.

# Contents

<b>Contents</b>	<b>13</b>
<b>List of Figures</b>	<b>17</b>
<b>List of Tables</b>	<b>19</b>
<b>1 Introduction</b>	<b>21</b>
1.1 Motivation . . . . .	21
1.1.1 Resource Classification . . . . .	23
1.1.2 Social Annotations . . . . .	23
1.2 Scope of the Thesis . . . . .	26
1.3 Problem Statement and Research Questions . . . . .	26
1.4 Research Methodology . . . . .	28
1.5 Structure of the Thesis . . . . .	29
1.6 Writing Conventions . . . . .	31
1.6.1 Formatting . . . . .	31
1.6.2 Language Issues . . . . .	31
<b>2 Related Work</b>	<b>33</b>
2.1 Resource Classification . . . . .	34
2.1.1 Binary and Multiclass Classification . . . . .	34
2.1.2 Single-label vs Multilabel Classification . . . . .	35
2.1.3 Semi-supervised vs Supervised Classification . . . . .	35
2.2 Support Vector Machines for Classification . . . . .	35
2.2.1 Semi-supervised Learning for SVM (S3VM) . . . . .	37
2.2.2 Multiclass SVM . . . . .	38
2.2.3 Multiclass S3VM . . . . .	39

2.3	Benefiting from Social Annotations . . . . .	40
2.3.1	Social Annotations for Information Management . . . . .	41
2.3.2	Social Annotations for Classification . . . . .	42
<b>3</b>	<b>Support Vector Machines for Large-Scale Classification</b>	<b>47</b>
3.1	Definition of Large-Scale Classification . . . . .	48
3.2	Compared SVM Approaches . . . . .	48
3.2.1	Native Multiclass Approaches . . . . .	49
3.2.2	One-Against-All Approaches . . . . .	50
3.2.3	One-Against-One Approaches . . . . .	50
3.3	Experiment Settings . . . . .	51
3.3.1	Datasets . . . . .	51
3.3.2	Document Representation . . . . .	52
3.3.3	Algorithmic Implementation . . . . .	52
3.3.4	Evaluation Measures . . . . .	53
3.4	Results . . . . .	53
3.4.1	Native Multiclass vs Combining Binary Classifiers . . . . .	53
3.4.2	Supervised vs Semi-Supervised Learning . . . . .	55
3.5	Discussion . . . . .	55
3.6	Conclusion . . . . .	56
<b>4</b>	<b>Generation of Social Tagging Datasets</b>	<b>59</b>
4.1	Selection of Social Tagging Systems . . . . .	60
4.2	Characteristics of the Selected Social Tagging Systems . . . . .	61
4.3	Generation Process of Datasets . . . . .	63
4.3.1	Getting Popular Resources . . . . .	63
4.3.2	Looking for Classification Data . . . . .	63
4.3.3	Gathering Tagging Data . . . . .	64
4.4	Statistics and Analysis of the Datasets . . . . .	65
4.5	Gathering Additional Data . . . . .	70
4.6	Conclusion . . . . .	72
<b>5</b>	<b>Representing the Aggregation of Tags</b>	<b>75</b>
5.1	Aggregation of User Annotations . . . . .	76
5.2	Representing Resources Using Tags . . . . .	77
5.3	Tag-based Classification . . . . .	79
5.4	Comparing Social Tags to Other Data Sources . . . . .	84
5.5	Getting the Most Out of All Data Sources . . . . .	86
5.6	Conclusion . . . . .	90

<b>6</b>	<b>Analyzing the Distribution of Tags for Resource Classification</b>	<b>95</b>
6.1	Tag Distributions . . . . .	96
6.2	TF-IDF as a Term Weighting Function . . . . .	97
6.3	Tag Weighting Functions Based on Inverse Frequencies . . . . .	98
6.3.1	TF-IRF . . . . .	99
6.3.2	TF-IUF . . . . .	100
6.3.3	TF-IBF . . . . .	100
6.4	Experiments . . . . .	100
6.4.1	Tag-based Classification . . . . .	101
6.4.2	Revisiting Classifier Committees . . . . .	104
6.4.3	Correlation between Tag Weighting Functions . . . . .	108
6.5	Conclusion . . . . .	109
<b>7</b>	<b>Analyzing the Behavior of Users for Classification</b>	<b>111</b>
7.1	User Behavior on Social Tagging Systems . . . . .	112
7.2	Categorizers vs Describers . . . . .	113
7.2.1	Measures . . . . .	113
7.2.1.1	Tags per Post (TPP) . . . . .	114
7.2.1.2	Orphan Ratio (ORPHAN) . . . . .	114
7.2.1.3	Tag Resource Ratio (TRR) . . . . .	114
7.3	Calculation of Measures and Experiment Settings . . . . .	115
7.3.1	Tag-based classification . . . . .	117
7.3.2	Descriptiveness of Tags . . . . .	117
7.4	Results . . . . .	119
7.4.1	Categorizers Perform Better on Classification . . . . .	119
7.4.2	Describers Perform Better on Descriptiveness . . . . .	122
7.4.3	Verbosity vs Diversity . . . . .	122
7.4.4	Non-descriptive Tags Provide More Accurate Classification . . . . .	123
7.5	Conclusion . . . . .	123
<b>8</b>	<b>Conclusions and Future Research</b>	<b>125</b>
8.1	Summary of Contributions . . . . .	125
8.2	Answers to Research Questions . . . . .	127
8.3	Future Directions . . . . .	131
	<b>Bibliography</b>	<b>133</b>
<b>A</b>	<b>Additional Results</b>	<b>143</b>
<b>B</b>	<b>Key Terms and Definitions</b>	<b>145</b>
<b>C</b>	<b>List of Acronyms</b>	<b>147</b>

<b>D Resumen (Spanish Summary)</b>	<b>149</b>
D.1 Motivación . . . . .	150
D.1.1 Clasificación de Recursos . . . . .	151
D.1.2 Anotaciones Sociales . . . . .	152
D.2 Objetivos . . . . .	155
D.3 Metodología . . . . .	155
D.4 Estructura de la Tesis . . . . .	156
D.5 Preguntas de Investigación Resueltas . . . . .	158
D.6 Principales Contribuciones . . . . .	163
D.7 Trabajo Futuro . . . . .	164
<b>E Laburpena (Basque Summary)</b>	<b>167</b>
E.1 Motibazioa . . . . .	167
E.1.1 Baliabideen Sailkapena . . . . .	169
E.1.2 Anotazio Sozialak . . . . .	170
E.2 Helburuak . . . . .	172
E.3 Metodologia . . . . .	173
E.4 Tesiaren Egitura . . . . .	174
E.5 Ebatzitako Ikerketa Galderak . . . . .	176
E.6 Ekarpen Nagusiak . . . . .	180
E.7 Etorkizunerako Ildoak . . . . .	181



## List of Figures

1.1	Simple tagging vs collaborative tagging . . . . .	25
2.1	Example of binary SVM classification . . . . .	36
2.2	Supervised vs Semi-Supervised SVM . . . . .	38
3.1	Splitting a training set into labeled and unlabeled subsets . . . . .	49
4.1	Tag usage percentages in the collection . . . . .	69
4.2	Tag usage percentages on resources . . . . .	70
4.3	Tag distribution across resources, users and bookmarks . . . . .	71
4.4	Novelty ratio of tags per rank of bookmark . . . . .	72
7.1	Example of splitting based on tag assignments or number of users	117
D.1	Etiquetado simple y etiquetado colaborativo . . . . .	154
E.1	Etiketatzte sinplea eta etiketatzen kolaboratiboa . . . . .	172



## List of Tables

3.1	Accuracy results for the BankSearch dataset . . . . .	54
3.2	Accuracy results for the WebKB dataset . . . . .	54
3.3	Accuracy results for the Yahoo! Science dataset . . . . .	54
4.1	Characteristics of the studied social tagging systems . . . . .	62
4.2	Number of resources and classes for the classification experiments	64
4.3	Statistics on availability of tags in users, bookmarks, and resources	66
4.4	Ratio of resources and bookmarks belonging to categorized or un-	
	categorized data . . . . .	67
4.5	Top 10 most popular tags . . . . .	67
4.6	Average counts of different tags . . . . .	68
5.1	Example of annotations for Flickr.com on Delicious . . . . .	76
5.2	Example of top 10 tags for Flickr.com on Delicious . . . . .	77
5.3	Example of the 4 representations of social tags . . . . .	79
5.4	Summary of tag representations . . . . .	79
5.5	Accuracy results for tag-based web page classification . . . . .	80
5.6	Accuracy results for tag-based book classification on LibraryThing	82
5.7	Accuracy results for tag-based book classification on GoodReads .	83
5.8	Accuracy results comparing different data sources on web page	
	classification . . . . .	84
5.9	Accuracy results comparing different data sources on book classi-	
	fication . . . . .	85
5.10	Example of classifier committees . . . . .	88
5.11	Accuracy results of classifier committees for web page classification	88

5.12	Accuracy results of classifier committees for book classification on LibraryThing . . . . .	89
5.13	Accuracy results of classifier committees for book classification on GoodReads . . . . .	91
6.1	Accuracy results of tag-based web page classification using weighting schemes . . . . .	101
6.2	Accuracy results of tag-based book classification using weighting schemes on LibraryThing . . . . .	102
6.3	Accuracy results of tag-based book classification using weighting schemes on GoodReads . . . . .	103
6.4	Accuracy results of classifier committees for web page classification using weighting schemes . . . . .	105
6.5	Accuracy results of classifier committees for book classification using weighting schemes on LibraryThing . . . . .	106
6.6	Accuracy results of classifier committees for book classification using weighting schemes on GoodReads . . . . .	107
6.7	Pearson and Spearman correlation coefficients . . . . .	108
7.1	Characteristics of Categorizers and Describers . . . . .	113
7.2	Measure distribution histograms . . . . .	116
7.3	Classification results for Categorizers and Describers . . . . .	120
7.4	Descriptiveness results for Categorizers and Describers . . . . .	121
A.1	Accuracy results of tag-based classification relying with different number of tags in the top. . . . .	144

*“Ideals are like stars; you will not succeed in touching them with your hands. But like the seafaring man on the desert of waters, you choose them as your guides, and following them you will reach your destiny.”*

— Carl Schurz

## 1.1 Motivation

Organizing resources into predefined categories is a natural idea in our daily lives. Assigning categories to resources helps facilitate the search for resources by reducing the focus to a specific category or categories. Categorization effectively reduces the amount of resources one has to search. For instance, librarians usually organize books into groups of related subjects. Also, movie databases, music catalogs, and file systems, among others, tend to be categorized in a way that eases access to their resources. Likewise, web directories such as the Yahoo! Directory and the Open Directory Project organize web pages into categories. Web page classification can substantially enhance search engines by reducing the scope of results to the category of user’s interest (Qi and Davison, 2009).

The process of manually categorizing resources becomes expensive as the collection of resources grows. For instance, the Library of Congress reported that the average cost of cataloging each bibliographic record by professionals was \$94.58 in 2002<sup>1</sup>. For the 291,749 records they cataloged that year, the total cost came to more than \$27.5 million. Given the expensiveness of this task, switching to automated classifiers seems to be a good alternative to facilitate the task and keep catalogs updated by reducing manual effort.

Until now, most of the automated classifiers rely on the content of the resources, especially regarding web page classification tasks (Qi and Davison (2009)).

---

<sup>1</sup><http://www.loc.gov/loc/lcib/0302/collections.html>

Nonetheless, the lack of representative data within many resources makes the classification task more complicated. In some cases, it may not be feasible to obtain enough data for certain kinds of resources such as books or movies. For example, usually the full text of books is not available, and it is not easy to represent movies as text or processable data. Without sufficient data, representing the content becomes more challenging.

As a means to solve these issues, social tagging systems provide an easier and cheaper way to obtain metadata related to resources. Social tagging systems are a means to save, organize, and search resources, by annotating them with tags that the user provides. Systems like Delicious<sup>2</sup>, LibraryThing<sup>3</sup> and GoodReads<sup>4</sup> collect user annotations in the form of tags on their respective collections of resources. These user-generated tags give rise to meaningful data describing the content of the resources (Heymann et al., 2008). User-provided annotations can be useful as a data source by providing meaningful information that can help infer the categorization of the resources. Our hypothesis is that these large collections of annotations can enhance the automated resource classification task in a noticeable manner.

By providing tags, users are creating their own categorization system for the given resource. The aggregation of users in an active community can create many bookmarks, tags, and therefore annotated resources. With more users contributing bookmarks and tags to these systems, the more accurately these resources can be annotated.

*“Each individual categorization scheme is worth less than a professional categorization scheme. But there are many, many more of them”,* Joshua Schachter, founder of Delicious, at the 2006 FOWA summit in London, England<sup>5</sup>.

Given that a large number of users are providing their own annotations on each resource, our objective is focused on finding out an approach to amalgamate their contributions in such a way that resembles the categorization by professionals. In this context, where users are providing large amounts of metadata, our challenge lies in making the most of them in order to enhance resource categorization tasks.

*“We’ve entered an era where data is cheap, but making sense of it is not”,* Danah Boyd, Social Media Researcher at Microsoft Research New England, at the WWW2010 conference in Raleigh, North Carolina,

---

<sup>2</sup><http://delicious.com>

<sup>3</sup><http://www.librarything.com>

<sup>4</sup><http://www.goodreads.com>

<sup>5</sup><http://simonwillison.net/2006/Feb/8/summit/>

United States<sup>6</sup>.

### 1.1.1 Resource Classification

Resource classification can be defined as the task of labeling and organizing resources within a set of predefined categories. In this work, we use Support Vector Machines (SVM, Joachims (1998)), a state-of-the-art classification approach. This type of classification relies on previously categorized or labeled training sets of resources. The classifier uses these sets of resources to gather knowledge which, in turn, is used to classify new unknown resources.

Different settings can be used for resource different classification problems. The system's learning technique may be *supervised* or *semi-supervised*. *Supervised* learning requires that all training resources are previously categorized where *semi-supervised* learning permits unlabeled resources to be taken into account during the learning phase. Classification may be *binary*, where only two possible categories can be assigned to each resource, or *multiclass*, where three or more categories can be assigned. *Binary* classification systems are commonly used for filtering systems –e.g., an email application that filters out spam messages–, whereas the *multiclass* systems are necessary for thematic classification with larger taxonomies –i.e., classification by topic or subject.

For thematic classification on large collections of resources, like web pages on the Web, or books in libraries, the taxonomies are usually defined by more than two categories, and the subset of previously labeled resources tends to be tiny. Accordingly, we believe that the application of both semi-supervised and multiclass approaches should be considered and analyzed to perform this kind of task.

In this thesis, we propose the analysis of several classification approaches using SVM, with the aim of analyzing their suitability to these tasks. These include different approaches to solving multiclass problems, as well as the study of *supervised* and *semi-supervised* algorithms.

### 1.1.2 Social Annotations

Social tagging sites allow users to save and annotate their favorite resources – e.g., web pages, movies, books, photos or music–, socially sharing them with the community. These annotations are usually provided by users in the form of tags. Tagging is an open way to assign tags or keywords to resources, in order to describe and organize them. It enables the later retrieval of resources in an easier way, using tags as metadata describing resources. Usually, there are no

---

<sup>6</sup><http://www.danah.org/papers/talks/2010/WWW2010.html>

predefined tags, and therefore users can freely choose the words they want as tags.

*“Tagging is mostly user interface - a way for people to recall things, what they were thinking about when they saved it. Fairly useful for recall, OK for discovery, terrible for distribution (where publishers add as many tags as possible to get it in lots of boxes).”*, Joshua Schachter, founder of Delicious, at the 2006 FOWA summit in London, England<sup>7</sup>.

This tagging process generates a tag structure so-called folksonomy on a social tagging system, i.e., a user-driven organization of resources. Folksonomy is a portmanteau of the words *folk* (people), *taxis* (classification) and *nomos* (management). It is also known as a community-based taxonomy, where the classification scheme is non-hierarchical, as opposed to a classical taxonomy-based categorization scheme. Thus, a folksonomy has to do with expert-driven taxonomies, insofar as resources are labeled and put together into groups.

These annotations are said to belong to a social environment when they are accessible and profitable by any user. This feature enables searching resources by taking advantage of annotations provided by others. This encourages the contribution of large communities of users.

Not all the annotations are shared in the same way, though. The social tagging site itself may establish some constraints, mainly by setting who is able to annotate each resource. In this regard, two kinds of systems can be distinguished (Smith, 2008):

- **Simple tagging systems:** users can describe their own resources, such as photos on Flickr<sup>8</sup>, news on Digg<sup>9</sup> or videos on Youtube<sup>10</sup>, but nobody annotates others' resources. Usually, the author of the resource is who annotates it. This means that no more than one user tags a resource. In a simple tagging system, there is a set of users ( $U$ ), who are annotating resources ( $R$ ) using tags ( $T$ ). A user  $u_i \in U$  annotates their resource  $r_j \in R$  with a set of tags  $T_j = \{t_{j1}, \dots, t_{jp}\}$ , with a variable number  $p$  of tags. The set of tags assigned to  $r_j$  will always be limited to  $T_j$ , since nobody else can annotate it.
- **Collaborative tagging systems:** many users annotate the same resource, and all of them can tag it with tags in their own vocabulary. The collection of tags assigned by a single user creates a smaller folksonomy, also

<sup>7</sup><http://simonwillison.net/2006/Feb/8/summit/>

<sup>8</sup><http://www.flickr.com>

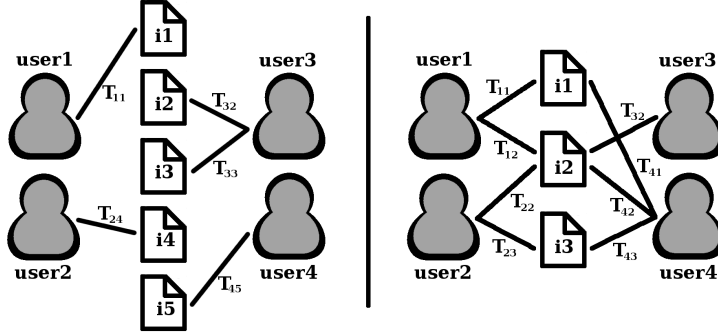
<sup>9</sup><http://digg.com>

<sup>10</sup><http://www.youtube.com>



known as personomy. As a result, several users tend to post the same resource. For instance, CiteULike<sup>11</sup>, LibraryThing and Delicious are based on collaborative annotations, where each resource (papers, books and URLs, respectively) can be annotated and tagged by all the users who consider it interesting. A collaborative tagging system is more complex than a simple one, where there is a set of users ( $U$ ), who are posting bookmarks ( $B$ ) for resources ( $R$ ) annotated by tags ( $T$ ). Each user  $u_i \in U$  can post a bookmark  $b_{ij} \in B$  of a resource  $r_j \in R$  with a set of tags  $T_{ij} = \{t_{ij1}, \dots, t_{ijp}\}$ , with a variable number  $p$  of tags. After  $k$  users posted  $r_j$ , it is described with a weighted set of tags  $T_j = \{w_{j1}t_{j1}, \dots, w_{jn}t_{jn}\}$ , where  $w_{j1}, \dots, w_{jn} \leq k$  represent the number of assignments of a specific tag. Accordingly, each bookmark is a triple of a user, a resource, and a set of tags:  $b_{ij} : u_i \times r_j \times T_{ij}$ . Thus, each user saves bookmarks of different resources, and a resource has bookmarks posted by different users. The result of aggregating tags within bookmarks by a user is known as the personomy of the user:  $T_i = \{w_{i1}t_{i1}, \dots, w_{im}t_{im}\}$ , where  $m$  is the number of different tags in user's personomy.

Figure 1.1 shows an example comparing the behavior of both systems.



**Figure 1.1:** Comparison of user annotations on simple and collaborative tagging systems.

In this thesis, we will focus on collaborative tagging systems. Tags present a high likelihood of coincidence across users annotating the same resource, making the aggregated tags of collaborative tagging systems especially strong rather than simple tagging systems, i.e., multiple users annotate the same resource.

In a collaborative tagging system, for instance, a user could tag this work as `social-tagging`, `research`, and `thesis`, whereas another user could use the tags `social-tagging`, `social-bookmarking`, `phd`, and `thesis` to annotate it. Users' behavior may considerably differ on these systems. Because of this, the

<sup>11</sup><http://www.citeulike.org>

aggregation of their annotations is usually considered as the consensus. For instance, the aggregation of the above annotations would be the following: *thesis* (2), *social-tagging* (2), *social-bookmarking* (1), *phd* (1), and *research* (1). In this example the values represent the weighted union of all tags.

In this thesis, we analyze and study the annotations provided by end users on social tagging systems. We present different methods to use these annotations to classify resources as accurately as possible. Specifically, we focus on the analysis of the usefulness of tags on user-driven folksonomies as a means to get an organization that resembles the categories on expert-driven taxonomies. In this context, we study several representations of social annotations, in search of an approach that resembles the classification by experts as much as possible. Especially, we focus on getting the most out of social tags, by both looking for the best representation, and measuring the impact of the distribution of tags across the triple of resources, bookmarks and users. Finally, we also study the application of state-of-the-art user behavior analysis approaches for the detection of users who rather provide tags for categorization purposes.

## 1.2 Scope of the Thesis

In this thesis, we investigate how annotations gathered together on social tagging systems can be harnessed for resource classification. Specifically, this thesis focuses on the study of several resource representation approaches using social tags. We perform the evaluation of such representations by measuring their similarity to classifications by experts. In this context, we consider the classification provided by experts as a ground truth for the evaluation process. We perform the classification experiments by using a state-of-the-art classification method, so-called Support Vector Machines. To choose the appropriate settings for the classifier, we also perform a preliminary study in this regard.

As meaningful metadata to enhance the resource classification task, we explore social tags provided by users from a statistical and distributional point of view, and we do not consider other details such as analyzing their linguistic and semantic meanings. For us, each text string representing a tag is treated as a different token, regardless of its meaning. Thereby, rather than analyzing the meaning of tags, we focus on analyzing the structure of folksonomies, represented by triples of users, bookmarks and resources.

## 1.3 Problem Statement and Research Questions

The main goal of this thesis is to shed new light on the appropriate use of the great deal of data gathered on social tagging systems. Given the interest of classifying

resources, and the lack of representative data in many cases, we aim at analyzing the extent to which and how social tags can enhance a resource classification task. At the beginning of this work, we found no works dealing with this insofar as no special attention had been paid at how to represent resources using social tags, and no actual classification experiments had been performed. Thus, we were motivated to carry out this research work. To that end, we set forth the following problem statement, which summarizes the main focus of this thesis:

**Problem Statement**

*How can the annotations provided by users on social tagging systems be exploited to yield the most accurate resource classification task?*

Regarding the classification algorithm, we rely on Support Vector Machines (SVM) as a state-of-the-art classification method. Using this method, several approaches have already been proposed to work on binary and multiclass scenarios, as well as supervised and semi-supervised ones. Nonetheless, there is little work comparing different approaches in the multiclass scenario. We assume that these kinds of tasks are usually multiclass, and the number of prior annotated resources tends to be tiny as compared to the whole collection of resources. Accordingly, the first two research questions we formulate in this thesis are:

**Research Question 1**

*What kind of SVM classifiers should be used to perform this kind of classification tasks: a native multiclass classifier, or a combination of binary classifiers?*

**Research Question 2**

*What kind of learning method performs better for this kind of classification tasks: a supervised one, or a semi-supervised one?*

Moreover, regarding social annotations, it has been shown that they provide useful metadata for improving resource management. Nevertheless, there is little work analyzing the usefulness of social tags for performing classification tasks. Preliminary analyses have shown encouraging results, and conclude that these annotations may be helpful for classification. However, they did not analyze the annotations in more depth, and it is not clear whether the representation they used was good enough.

We believe that several factors should be taken into account when representing resources using social tags. In contrast to classical document repositories, social annotations rely on a triple of users, resources, and tags, which should be analyzed in more depth for the representation task. In this context, apart from representing the resources, it is worthwhile considering that not all the tags have to be equally representative, and not all the users provide equally good annotations. In this thesis, our main goal is to deepen on the way social annotations

can be used to the greatest extent in search of an accurate classification of the resources. Based on these ideas, we formulate the following research questions:

**Research Question 3**

*How do the settings of social tagging systems affect users' annotations and the resulting folksonomies?*

**Research Question 4**

*What is the best way of amalgamating users' aggregated annotations on a resource in order to get a single representation for a resource classification task?*

**Research Question 5**

*Despite of the usefulness of social tags for these tasks, is it worthwhile considering their combination with other data sources like the content of the resource as an approach to improve the results even more?*

**Research Question 6**

*Are social tags also useful and specific enough to classify resources into narrower categories as in deeper levels of hierarchical taxonomies?*

**Research Question 7**

*Can we further consider the distribution of tags across the collection so that we can measure the overall representativity of each tag to represent resources?*

**Research Question 8**

*What is the best approach to weigh the representativity of tags in the collection for resource classification?*

**Research Question 9**

*Can we discriminate different user profiles so that we can find a subset of users who provide annotations that better fit a classification scheme?*

**Research Question 10**

*What are the features that identify a user as a good contributor to the resource classification?*

## 1.4 Research Methodology

The research methodology we followed throughout this work includes 6 parts:

1. Review of the literature and understanding of social tagging systems.
2. Looking for an appropriate SVM classifier to perform the work.
3. Looking for existing social tagging datasets. Since we did not find any that fulfilled our requirements, we created three large-scale social tagging datasets instead.

4. Thinking of and proposing approaches to classifying using social tags.
5. Evaluating the proposed approaches.
6. Performing a thorough analysis of the results, in order to understand them for drawing conclusions.
7. Showing and presenting partial results at several national and international conferences and workshops, in order to get useful comments and feedback from other researchers.
8. Summarizing the research, contributions, and conclusions drawn throughout this work by writing this dissertation.

Step 4 through 6 was an iterative process.

## 1.5 Structure of the Thesis

This thesis consists of 8 chapters. Below we provide a brief overview summarizing the contents of each of these chapters.

### Chapter 1 on page 21

#### Introduction

We present the motivation for the study on the use of social annotations for resource classification. We formalize the problem, and motivate the need of such a study.

### Chapter 2 on page 33

#### Related Work

We provide a survey of previous works in the field. We summarize the advances in related fields, not only on the use of social annotations, but also on resource classification.

### Chapter 3 on page 47

#### Support Vector Machines for Large-Scale Classification

We perform a study on different SVM approaches to the problem of classifying large-scale resource collections on multiclass taxonomies. It gives rise to the best SVM approach, which we use to perform the rest of the classification experiments along the work.

### Chapter 4 on page 59

#### Generation of Social Tagging Datasets

We describe and analyze in detail the social tagging datasets we created. We detail in depth the process of creation of such datasets, and we analyze the main characteristics of the underlying folksonomies.

**Chapter 5 on page 75****Representing the Aggregation of Tags**

We propose and evaluate different representations of resources based on social tags for a resource classification task. We study the usefulness of social tags as compared to other data sources, and propose the best representation approach to get the most out of them. We also deal with the combination of social tags with other data sources to yield a better performance.

**Chapter 6 on page 95****Analyzing the Distribution of Tags for Resource Classification**

We deal with the task of considering the representativity of tags within a collection of social annotations on a social tagging system for resource classification. We study the application of weighting schemes adapted to social tagging systems, and analyze their suitability by taking into account the settings of each system.

**Chapter 7 on page 111****Analyzing the Behavior of Users for Classification**

We explore the effect of user behavior on social tagging systems for the resource classification task. Previous works suggest the existence of two types of users: Categorizers, who use tags to categorize resources, and Describers, who use tags to describe resources. Based on these works, we study whether tags by Categorizers are better than tags by Describers for the resource classification.

**Chapter 8 on page 125****Conclusions and Future Research**

We discuss and summarize the main conclusions and contributions of the work. We present the answers to the formulated research questions, and the outlook on future directions of the work.

Additionally, the thesis contains the following appendices at the end, with complementary information and summaries in other languages:

**Appendix A on page 143****Additional Results**

We present some additional results, which we did not include in the main content of the thesis, but are also worth including to prove and help understand some conclusions.

**Appendix B on page 145****Key Terms and Definitions**

We list the most relevant terms related to social tagging systems, and provide a detailed definition of them.

**Appendix C on page 147****List of Acronyms**

We provide a list of the acronyms used along the work, and what they stand for.

**Appendix D on page 149****Resumen (Spanish Summary)**

We summarize the contents of this work in Spanish language.

**Appendix E on page 167****Laburpena (Basque Summary)**

We summarize the contents of this work in Basque language.

## 1.6 Writing Conventions

Next, we detail some conventions we defined while writing this thesis. These conventions include formatting of text, and some issues regarding English language.

### 1.6.1 Formatting

In the thesis, we mention names of tags many times, either to show them as examples or to clarify some explanations. When those tags appear in the text, we use a monospaced typeface to differentiate them easily from the rest of the text. For instance: `reference`.

In the same manner, we emphasize with italic text those inline appearances of math formulas, or terms that for some reason have certain importance in the context.

### 1.6.2 Language Issues

This thesis, being focused on social media, deals with users of social tagging systems at some points. When we refer to a single user, but no distinction is made between genders, we use the pronoun *they* instead of either *he* or *she*. For instance:

*When a user saves a bookmark, the tags annotated by **them** are added to **their** personomy.*

This is grammatically incorrect in English. However, a person's gender is explicit in the third person singular pronouns, and there is no perfect solution to this issue. Sometimes, the wording *he/she* is used, but using it all along this work would become cumbersome, and would harm its readability. We rely on tips by the Oxford English Dictionary for this decision<sup>12</sup>.

---

<sup>12</sup><http://www.oxforddictionaries.com/page/heshethey/he-or-she-versus-they>





## Related Work

*“If you have an apple and I have an apple and we exchange these apples then you and I will still each have one apple. But if you have an idea and I have an idea and we exchange these ideas, then each of us will have two ideas.”*

— George Bernard Shaw

This chapter introduces the previous work we found in the literature. Specifically, the works in the research areas related to this work are put together, summarized, and contextualized. Next, in [Section 2.1 on the next page](#) we define and provide a background on the resource classification problem. In [Section 2.2 on page 35](#) we summarize the previous efforts towards an SVM approach that enables the classification of resources within an environment where the taxonomy is multiclass (i.e., made up by more than two classes), and the number of labeled resources use to be tiny as compared to the unlabeled ones. In [Section 2.3 on page 40](#) we summarize the works in which annotations from social tagging systems have been profited to enhance information search, management and access. Specifically, we first summarize in [Subsection 2.3.1 on page 41](#) the use of social annotations to enhance information management tasks. Then, we present in detail in [Subsection 2.3.2 on page 42](#) the works regarding the use of social annotations for classification. The latter is the most important topic for this thesis, but due to the novelty of the research field and the lack of work on it, it does not extend as much as the former.

## 2.1 Resource Classification

Resource classification is the task of assigning categories from a predefined taxonomy to a set of resources. Formally, it consists of associating a Boolean value to each pair  $\langle r_j, c_i \rangle \in R \times C$ , where  $R = \{r_1, \dots, r_{|R|}\}$  is the set of resources, and  $C = \{c_1, \dots, c_{|C|}\}$  is the set of predefined categories. The goal of the task aims at letting the classifier give predictions by means of the function  $\phi^* : R \times C \rightarrow \{T, F\}$ , in such a way that it resembles as much as possible the function  $\phi : R \times C \rightarrow \{T, F\}$ , which defines the ideal classification of the resources (Sebastiani, 2002). Upon this, several settings can define a different classification approach. Next, we briefly define the main settings.

Usually, a classification task comprises two subsets of resources when it relies on a machine learning approach. Some of the resources are already labeled with corresponding categories, and others are unlabeled. The former are used by the classifier to learn the characteristics of each category, creating a model for each category after the learning process. The latter are the instances to be predicted by the classifier. Relying on the models created during the learning phase, the classifier provides a category for each unlabeled resource as its prediction.

### 2.1.1 Binary and Multiclass Classification

As regards to the taxonomies with predefined categories considered in the classification, where the resources are organized, the task is said to be either binary or multiclass. Even though the sole apparent difference is the number of classes making up the taxonomy –2 for the binary, and 3 or more for the multiclass–, the tasks tend to follow a different goal.

A binary classification is usually part of a filtering process, where the classes are the positive and the negative case. These tasks aim at separating the resources that want to be considered from the resources that want to be ruled out. For instance, a common binary classifier used as a filter is an email application that keeps the interesting messages in the inbox, whereas it sends the unwanted stuff to the spam folder.

A multiclass classification involves a larger taxonomy, and is usually used on thematic classification, i.e., where the categories represent the aboutness of the underlying resources. This kind of classification enables to organize resources into groups of related matters, and it has several applications such as creating directories of resources to ease later browsing, providing customized suggestions by users' topics of interest, or allowing to handle resources from different categories in a separate way, among many others.

In this thesis, we deal with thematic classification and, thus, we consider the task to be multiclass.

### 2.1.2 Single-label vs Multilabel Classification

The number of categories or labels that can be assigned to each resource is another setting that describes a resource classification. The number of labels for a single resource can be constrained to just one, thus becoming a single-label task, or it must be extended to allow more labels or even unlimited, when it is called multilabel. In practice, it refers to whether a resource can be related to several categories, or it can be included into just one. Besides the classification task itself, this feature also modifies the subsequent organization and browsing of categories.

In this thesis, we focus on single-label classification, mainly because the taxonomies we use as the ground truth provide this kind of categorization data.

### 2.1.3 Semi-supervised vs Supervised Classification

With regard to the learning method used by the classifier, it can vary in the instances considered to learn and create the model. A supervised learning method learns from the instances in the training set, and creates a model from them. A semi-supervised goes further by also considering unlabeled instances in the learning method. After creating the model from labeled instances, it includes its predictions on the unlabeled instances in the learning process enabling an incremental evolution of the model. The latter is especially useful when the training set is small, and the lack of sufficient learning data is worth an upsize of the labeled data.

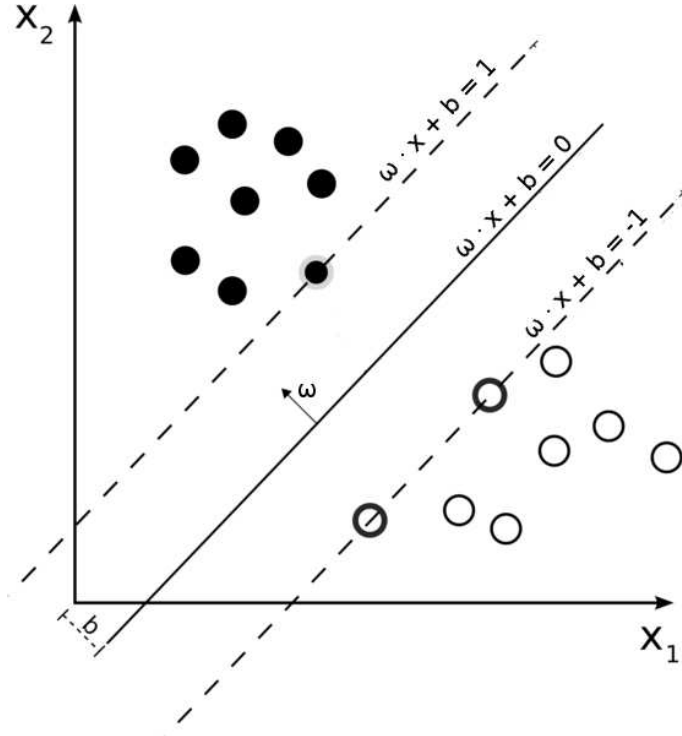
Based upon these two learning methods, we summarize the related work on the use of SVM classifiers in the next section. We rely on SVM as a state-of-the-art classification algorithm widely used in the field.

## 2.2 Support Vector Machines for Classification

In the last decade, SVM has become one of the most widely studied techniques for text classification, due to the positive results it has shown. This technique uses the vector space model to represent the resources, and assumes that resources in the same class should fall into separable spaces of the representation. Upon this, it looks for a hyperplane that separates the classes; therefore, this hyperplane should maximize the distance between it and the nearest resources, which is called the margin. Equation 2.1 defines such a hyperplane (see Figure 2.1 on the following page).

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b \quad (2.1)$$

In order to resolve this function, though, all the possible values should be considered and, after that, the values of  $w$  and  $b$  that maximize the margin should



**Figure 2.1:** An example of binary SVM classification, separating two classes (black dots from white dots). Source: Wikimedia Commons.

be selected; this would be computationally expensive. The equivalent Equation 2.2 is thus used to relax it (Boser et al., 1992; Cortes and Vapnik, 1995):

$$\min \left[ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^l \zeta_i^d \right] \quad (2.2)$$

Subject to:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \zeta_i, \quad \zeta_i \geq 0$$

where  $C$  is the penalty parameter,  $\zeta_i$  is a slack variable for the  $i^{th}$  resource,  $l$  is the number of labeled resources, and  $d$  is the sigma parameter which defines the non-linear mapping from the input space to some high-dimensional feature space.

When the value of  $d$  is set to 1, this function can only solve linearly separable problems. The use of a kernel function is sometimes required for the redimension of the space. This redimension creates a new space with higher number of

dimensions, which enables a linear separation. After that, the redimension is undone, so the hyperplane will be transformed to the original space, respecting the classification function. Best-known kernel functions include linear, polynomial, radial basis function (RBF) and sigmoid, among others. Different kernel functions' performance has been studied in [Schölkopf et al. \(1999\)](#) and [Kivinen and Williamson \(2002\)](#). Linear kernel is most widely used for text classification.

Note that the function above can only resolve binary and supervised problems, so different variants are necessary to handle semi-supervised or multiclass tasks.

### 2.2.1 Semi-supervised Learning for SVM ( $S^3VM$ )

Semi-supervised learning approaches differ in the learning way of the classifier. As opposed to supervised approaches, unlabeled data is used during the learning phase. Taking into account unlabeled data to learn can help improve the performance of supervised classifiers, especially when its predictions provide new useful information, as shown in Figure 2.2. However, the noise added by incorrect predictions can worsen the learned model and, therefore, the performance of the classifier. This makes interesting the study on whether relying on semi-supervised approaches is suitable for a certain kind of task.

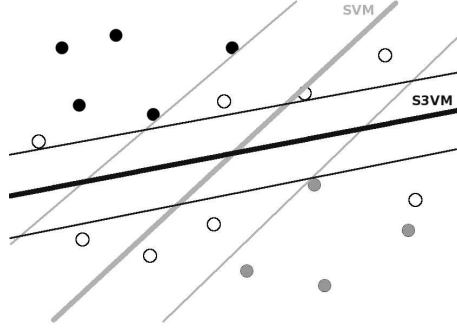
Semi-supervised learning for SVM, also known as  $S^3VM$ , was first introduced by [Joachims \(1999\)](#) in a transductive way, by modifying the original SVM function. To do that, the author proposed to add an additional term to the optimization function (see Equation 2.3).

$$\min \left[ \frac{1}{2} \cdot \|\mathbf{w}\|^2 + C \cdot \sum_{i=1}^l \xi_i^d + C^* \cdot \sum_{j=1}^u \xi_j^{*d} \right] \quad (2.3)$$

where  $u$  is the number of unlabeled data, and the parameters with an asterisk (\*) refer to the unlabeled instances included in the learning phase.

Nevertheless, the adaptation of SVM to semi-supervised learning significantly increases its computational cost, due to the non-convex nature of the resulting function, and so obtaining the minimum value is even more complicated. In order to relax the function, convex optimization techniques such as semi-definite programming are commonly used ([Xu et al., 2008](#)), where minimizing the function gets much easier.

By means of this approach, [Joachims \(1999\)](#) demonstrated a large performance gap between the original supervised SVM and his proposal for a semi-supervised SVM, in favor of the latter one. He showed that for binary classification tasks, the smaller is the training set size, the larger gets the difference among these two approaches. He used the Reuters-21578, Ohsumed and WebKB datasets for that purpose. Although he worked with multiclass datasets, he split



**Figure 2.2:** SVM vs S<sup>3</sup>VM, where black and grey dots are labeled resources, and white dots are unlabeled resources. It can be seen that the few labeled resources give rise to a certain separation (grey line, SVM), whereas including unlabeled ones helps infer a more accurate separation (black line, S<sup>3</sup>VM).

the problems into smaller binary ones, and so he did not demonstrate whether the same performance gap occurs for multiclass classification. More recently, [Chapelle et al. \(2008\)](#) presented a comprehensive review of advances in binary S<sup>3</sup>VM approaches.

### 2.2.2 Multiclass SVM

Due to the dichotomic nature of SVM, it came up the need to implement new methods to solve multiclass problems, where more than two classes must be considered. Different approaches have been proposed to achieve this. On the one hand, as a native approach, [Weston and Watkins \(1999\)](#) proposed modifying the optimization function getting into account all the  $k$  classes at once (see Equation 2.4).

$$\min \left[ \frac{1}{2} \sum_{m=1}^k \|\mathbf{w}_m\|^2 + C \sum_{i=1}^l \sum_{m \neq y_i} \xi_i^m \right] \quad (2.4)$$

Subject to:

$$\mathbf{w}_{y_i} \cdot \mathbf{x}_i + b_{y_i} \geq \mathbf{w}_m \cdot \mathbf{x}_i + b_m + 2 - \xi_i^m, \xi_i^m \geq 0$$

The main novelty of this approach, as compared to previous ones, is that apart from the choice of a kernel, it is parameterless. Their experiments on benchmark datasets from the UCI repository show results similar to SVMs which have been tuned to have the best choice of parameter.

On the other hand, the original binary SVM classifier has usually been combined to obtain a multiclass solution. As combinations of binary SVM classifiers,

two different approaches to  $k$ -class classifiers can be emphasized (Hsu and Lin, 2002):

- *one-against-all* constructs  $k$  classifiers defining that many hyperplanes, each of them separating the class  $i$  from the rest  $k-1$ . For instance, for a problem with 4 classes, the following classifiers would be created: 1 vs 2-3-4, 2 vs 1-3-4, 3 vs 1-2-4 and 4 vs 1-2-3. Unlabeled resources will be categorized in the class of the classifier that maximizes the margin:  $\hat{C}_i = \arg \max_{i=1,\dots,k} (w_i x + b_i)$ . As the number of classes increases, the amount of classifiers will increase linearly.
- *one-against-one* constructs  $\frac{k(k-1)}{2}$  classifiers, one for each possible category pair. For instance, for a problem with 4 classes, the following classifiers would be created: 1 vs 2, 1 vs 3, 1 vs 4, 2 vs 3, 2 vs 4 and 3 vs 4. After that, it classifies each new document by using all the classifiers, where a vote is added for the winning class over each classifier; the method will propose the class with more votes as the result. As the number of classes increases, the amount of classifiers will increase in an exponential way, and so the problem could become very expensive for large taxonomies.

Both Weston and Watkins (1999) and Hsu and Lin (2002) compare the native multiclass approach to the *one-against-one* and *one-against-all* binary classifier combining approaches. They agree concluding that the native approach does not outperform the results by *one-against-one* or *one-against-all*, although it considerably reduces the computational cost because the number of support vector machines it defines is smaller. Among the binary combining approaches, they show the performance of *one-against-one* to be superior to *one-against-all*.

Although these approaches have been widely used in supervised learning environments, they have scarcely been applied to semi-supervised learning. Accordingly, we believe that the study on its applicability and performance for this type of problems is necessary before proceeding with additional experiments.

### 2.2.3 Multiclass $S^3VM$

When the taxonomy is defined by more than two classes and the number of previously labeled documents is very small, the combination of both multiclass and semi-supervised approaches could be required, i.e., a multiclass  $S^3VM$  approach. A common web page classification problem meets these characteristics, with a taxonomy of more than two categories, and it could be helpful to increase the tiny amount of labeled documents by including predictions on unlabeled data for the learning phase.

However, little work has been done on transforming SVM into both a semi-supervised and multiclass approach, and especially on comparing them to other

approaches. As a native approach, [Yajima and Kuo \(2006\)](#) modified the original SVM function by fitting it to multiclass semi-supervised tasks (see Equation 2.5).

$$\min \frac{1}{2} \sum_{i=1}^h \beta^{i^T} K^{-1} \beta^i + C \sum_{j=1}^l \sum_{i \neq y_j} \max\{0, 1 - (\beta_j^{y_j} - \beta_j^i)\}^2 \quad (2.5)$$

where  $\beta$  represents the product of a vector of variables and a kernel matrix defined by the author.

The authors showed that the proposed approach outperformed other non-SVM algorithms, but they did not show if it was better than other SVM settings. As far as we know, the software was not made publicly available, and no further work has been done using this approach.

[Chapelle et al. \(2006\)](#) present another native multiclass S<sup>3</sup>VM approach by using the Continuation Method. This is the only work, to the best of our knowledge, where *one-against-all* and *one-against-one* approaches had been tested in a semi-supervised environment. They apply these methods to news datasets, yielding worse performance. Moreover, they show that *one-against-one* is not sufficient for real-world multiclass semi-supervised learning, since the unlabeled data cannot be restricted to the two classes under consideration.

On the other hand, others relied on combining SVM with other algorithms in search of a multiclass semi-supervised SVM approach. [Qi et al. \(2004\)](#) use Fuzzy C-Means (FCM) to predict labels on unlabeled resources. After that, multiclass SVM is used to learn with the augmented training set, classifying the test set. [Xu and Schuurmans \(2005\)](#) rely on a clustering-based approach to label the unlabeled data. Afterwards, they apply a multiclass SVM classifier to the labeled training set.

It is worthwhile noting that most of the above works introduced their approaches and only compared them to other semi-supervised classification methods, such as Expectation-Maximization (EM) or Naive Bayes. As an exception, [Chapelle et al. \(2006\)](#) compared a semi-supervised and a supervised SVM approach, but only over image datasets. In this thesis, we do not aim at proposing new SVM approaches. However, we believe that evaluating and comparing multiclass SVM and multiclass S<sup>3</sup>VM approaches is necessary to conclude with a suitable approach. This would help discover whether learning upon unlabeled resources is helpful for multiclass problems when using SVM as a classifier.

## 2.3 Benefiting from Social Annotations

Since it was introduced along with the Web 2.0 phenomenon in the early 2000s, social annotations have gained popularity and interest with the creation of well-known social tagging sites like Delicious. This section summarizes how social



annotations have helped improve information access and management. Among the existing works, it is especially focused on their use for classification tasks.

Social tagging systems arose as an idea of Joshua Schachter, founder of Delicious ([Smith, 2008](#)). In late 1990s, after the bookmarks in his web browser had overflowed, he used to save his favorite URLs in a text file, with an entry per line. Each entry was a URL, followed by a set of tags. These tags enabled him to easily refine the URL he was looking for. He just had to filter by keyword to search for a URL. He also published online such a list at Muxway.org (currently discontinued and not accessible). Later, in September 2003, he released Delicious, the first online tool that enabled saving and tagging URLs as he used to, but enhanced by a social environment. This social tool enabled users to search among saved URLs, not only by their own tags, but also taking advantage of others'.

The research on social tagging systems did not arise until 2006. An early work by [Golder and Huberman \(2006\)](#) performed a study of the characteristics of Delicious, followed by an increasingly interest of researchers that gave rise to large number of research works in the field. Next, we focus on some of the most relevant advances on the use of social tags for information management, and go in more depth for the specific task of resource classification.

### 2.3.1 Social Annotations for Information Management

Social annotations have been widely used for the sake of information management tasks. They have shown to be very useful for several tasks in which the availability of data is of utmost importance ([Gupta et al., 2010](#)):

- **Search:** Social tags have been successfully applied to web search. [Bao et al. \(2007\)](#) found that social annotations can enhance web search (1) as a good summaries of corresponding web pages, and (2) as a way to compute the popularity of web pages by considering the number of users who annotate them. [Heymann et al. \(2008\)](#) analyzed the usefulness of tags from Delicious for web search, and concluded that these metadata can provide additional and meaningful data not available in other sources, though it may currently lack the size to get a significant impact. Also, [Dmitriev et al. \(2006\)](#) showed the usefulness of social annotations for improving the quality of intranet search. As a specific approach for searching on social tagging systems, [Hotho et al. \(2006\)](#) proposed FolkRank, a search algorithm that fits the structure of folksonomies. They found this approach useful for providing personalized rankings of the resources in a folksonomy, as well as for finding communities of users within these systems.
- **Recommender Systems:** [Shepitsen et al. \(2008\)](#) and [Li et al. \(2008\)](#) introduced recommendation algorithms based on user-generated tags. They show that social annotations are effective to discover user interests and,

therefore, to recommend them new resources. In [Bogers and van den Bosch \(2008\)](#), the authors take advantage of annotations provided by users on a social reference manager for recommending research papers to scientists. In [Cantador et al. \(2011\)](#), the authors present a mechanism to automatically filter and classify social tags in a set of purpose-oriented categories, so that they can rely on suitable tags to recommend resources to users.

- **Enhanced Browsing:** Social annotations can be helpful to improve the navigation of resources as well. [Smith \(2008\)](#) describes three new navigation ways emerged from folksonomies: pivot browsing (moving through an information space by choosing a reference point to browse), popularity-driven navigation (retrieving the resources that are popular for a given tag), and filtering (social tagging allows to separate the resources you do not want from the resources you do want). In a preliminary study, we integrated tags from Delicious to Wikipedia ([Zubiaga, 2009](#)). Tags provided new data that was not available in the content of encyclopedia articles, providing a means to enhance the navigation and search on the site.
- **Clustering and Classification:** Social tags have also shown to be useful for resource organization tasks, including clustering and classification. This point will be explained in more depth in the next section.

### 2.3.2 Social Annotations for Classification

Before we began working on this thesis, there was little work dealing with the analysis of the applicability and usefulness of social tags for resource classification tasks. Most of them had focused on classifying web pages, and had just explored the appropriateness of social tags for this kind of tasks. However, none of them performed real classification experiments but just statistical analyses. Accordingly, they did not further explore on how to get the most out social tags in order to improve the performance.

[Noll and Meinel \(2008a\)](#) presented a study of the characteristics of social annotations provided by end users, in order to determine their usefulness for web page classification. In this work, the authors weight the tags by normalizing the number of users annotating them. The least popular tag is given a value of 0, whereas the most popular is given a value of 1. This way, they remove those least popular tags as they were useless. Moreover, they did not pay attention at whether or not this representation approach was appropriate to carry out the task. The authors matched user-supplied tags of a page against its categorization by the expert editors of the Open Directory Project (ODP). They analyzed at which hierarchy depth matches occurred, concluding that tags may perform better for broad categorization of documents rather than for more specific categorization. The study also points out that since users tend to bookmark and

tag top level web documents, this type of metadata will target classification of the entry pages of websites, whereas classification of deeper pages might require more direct content analysis. They observed that in the power law curve formed by the popularity of social tags, not only popular tags, but also the tags in the tail provide helpful data for information retrieval and classification tasks in general. In a previous work, the same authors (Noll and Meinel, 2007) suggested that tags provide additional information about a web page, which is not directly contained within its content.

Also, Noll and Meinel (2008b) studied three types of metadata about web documents: social annotations (tags), anchor texts of incoming hyperlinks, and search queries to access them. They concluded that tags are better suited for classification purposes than anchor texts or search keywords.

As regards to clustering tasks, Ramage et al. (2009) included tagging data for improving the performance of two clustering algorithms when compared to content-based clustering. They found that tagging data was more effective for specific collections than for a collection of general documents. They weighted the tags by both using the number of users annotating them, and reweighting this value considering the Inverse Document Frequency (IDF) value of the tag across the resources in the collection. They showed a superiority for the use of the IDF weighting scheme.

Even though those were the only works published by then, the interest on this research area has increased lately. After the presentation of our earliest work in the field (Zubiaga et al., 2009d), more scientists have shown their interest on it, and have presented new works.

In Aliakbary et al. (2009), the authors integrated social annotations as an approach to extending web directories. They relied on the number of users annotating each tag as a weight. Upon that, they created a model vector for each category, and computed the cosine similarity to new web pages to generate predictions. They observed that the annotations provided a multi-faceted summary of the web pages, and that they better represent the aboutness of web pages than the content itself. Also, they conclude that the more users annotate a URL, the better it is classified.

In another work where social tags were exploited for the benefit of web page classification, Godoy and Amandi (2010) also showed the usefulness of social tags for web page classification, which outperformed classifiers based on full-text of documents. Similar to our previous work (Zubiaga et al., 2009d), they compare tag-based resource representations relying on all the tags and the top 10 of tags for each resource, corroborating our findings that the former performs better. Going further, they concluded that stemming the tags reduces the performance of such classification, even though some operations such as removal of symbols, compound words and reduction of morphological variants have a discrete posi-

tive impact on the task.

Xia et al. (2010) studied the usefulness of social tags as a complementary source for improving the classification of academic conferences into corresponding topics. Using tagging data gathered from WikiCFP<sup>1</sup>, and weighing the tags according to the number of users annotating them, they compare the classification of conferences by using only the content of the call for papers, and by integrating tagging data along with it. Their experiments yielded slightly better performance for the integration of social tags, with roughly 1% improvement.

With regard to the classification of resources other than web pages, Lu et al. (2010) present a comparison of tags annotated on books and their Library of Congress subject headings. Actually, no classification experiments are performed, but a statistical analysis of the tagging data shows encouraging results. By means of a shallow analysis of the distribution of tags across the subject headings, they conclude that user-generated tags seem to provide an opportunity for libraries to enhance the access to their resources.

Using a graph-based approach, in Yin et al. (2009) the authors present a method to classify products from Amazon into their corresponding categories using social tags. They conclude that social tags can enhance web products classification by representing them in a meaningful feature space, interconnecting them to indicate relationship, and bridging heterogeneous products so that category information can be propagated from one domain to another.

There is also a set of works dealing with user behavior in social tagging systems. Even though they do not perform classification experiments, they suggest the existence of a subset of users in these systems who are rather categorizing the resources when annotating. Specifically, early works such as Hammond et al. (2005b) and Marlow et al. (2006a) suggest the existence of two types of users: on one hand, users can be motivated by categorization (so-called *Categorizers*). These users view tagging as a means to categorize resources according to some (shared or personal) high-level conceptualizations. On the other hand, users who are motivated by description (so-called *Describers*) view tagging as a means to accurately and precisely define the content of resources. These proposals have been further studied in Körner et al. (2010a) and Körner et al. (2010b). However, they focused on showing that the difference of motivation among those two kinds of users actually exists, and they did not pay attention at whether *Categorizers* are better suited to the classification task.

Also, there has recently been an increasingly interest in using social tags for the benefit of clustering tasks. In Lu et al. (2009) the authors not only cluster the annotated resources, but also users and tags. Zhang et al. (2009) found that the effectiveness of clustering blog posts using tags from a simple tagging system was quite limited, and they combined these data with relations in the blogosphere to

---

<sup>1</sup><http://www.wikicfp.com>

get better results.

So far, there is little work on the analysis of folksonomies for the classification task. A few works have shown their suitability for this purpose, but no special attention has been paid into further studying these metadata structures. In this thesis, we aim at analyzing these structures to find an approach to amalgamate and represent the tags in order to perform a resource classification task with a high accuracy.



## Support Vector Machines for Large-Scale Classification

*“Science is the systematic classification of experience.”*

— George Henry Lewes

We study the appropriateness of several SVM approaches for large-scale classification in this chapter. We are not going to introduce any new approaches in it, but to perform a preliminary comparison study among different approaches, in search of a suitable approach for large-scale classification tasks. We also evaluate the real contribution of unlabeled data for multiclass SVM-based classification tasks. Specifically, we compare a native multiclass approach to the combination of binary classifiers, as well as a supervised to a semi-supervised approach. We carry out such an experimentation by using three web page datasets. The Web is a good example of the problem we are dealing with, where the number of resources is very large, and the number of labeled ones tends to be tiny as compared to the whole collection. The chapter is organized as follows. Next, in [Section 3.1 on the next page](#) we define and present the features of a large-scale classification task. We enumerate and describe the SVM approaches compared in our study in [Section 3.2 on the following page](#). Then, in [Section 3.3 on page 51](#) we detail the settings of the experiments, showing their results in [Section 3.4 on page 53](#). Finally, we discuss the results in [Section 3.5 on page 55](#), and conclude and answer the following research questions in [Section 3.6 on page 56](#):

**Research Question 1**

*What kind of SVM classifiers should be used to perform this kind of classification tasks: a native multiclass classifier, or a combination of binary classifiers?*

**Research Question 2**

*What kind of learning method performs better for this kind of classification tasks: a supervised one, or a semi-supervised one?*

### 3.1 Definition of Large-Scale Classification

The Web comprises lots of collections of web documents and other resources that scale up constantly. The increasingly amount of resources on the Web has in part influenced the recent upsize of research on large-scale datasets. Accordingly, extending earlier studies on SVM classification to large-scale collections rises in importance.

In this regard, we believe that a thematic classification task commonly meets the following conditions when it comes to large-scale collections of resources:

- **Tininess of the training set:** getting a manual classification of a subset as a training set is very expensive and entails a lot of time and effort. Thus, the previously categorized subset will be tiny as compared to the uncategorized subset. This suggests considering semi-supervised approaches besides supervised ones, as a way of extending the training set.
- **Multiclass taxonomy:** a taxonomy is usually composed of more than two categories, and thereby it is considered as a multiclass task instead of a binary one.

We assume that the large-scale thematic classification tasks we are undertaking in this thesis will fulfill those two features.

### 3.2 Compared SVM Approaches

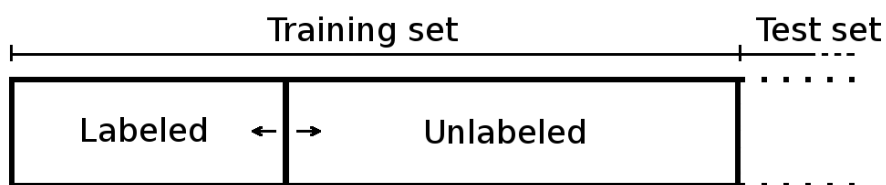
The two features showed above encourage the study of different SVM settings to conclude with a suitable one. On the one hand, the tininess of the training set requires analyzing whether or not it is worthwhile relying on a semi-supervised approach extending it instead of a supervised one. An early work by [Joachims \(1999\)](#) showed the outperformance of the former for binary classification, but it is not clear whether or not the same happens for multiclass scenarios. Especially because labeling instances on a larger taxonomy is harder, and it seems much



likelier to introduce noise in the extension of the training set. On the other hand, the classification on a multiclass taxonomy can be faced up in two different ways: as a single multiclass task, or as several smaller binary tasks. Little work has been done comparing these two settings, though. The lack of analyses on the appropriateness of the aforementioned methods for the task conveyed us to perform such a study prior to getting to work on large-scale classification tasks.

Our study involves both supervised SVM and semi-supervised SVM ( $S^3VM$ ) approaches on different multiclass settings. Specifically, we rely on three multiclass settings which were introduced in Section 2.2.2 on page 38: a native multiclass algorithm, and *one-against-one* and *one-against-all* based on binary classifiers. When naming the approaches, we add a suffix *-mSVM*, *-SVM* or *- $S^3VM$*  for clarity, depending if they are multiclass, supervised or semi-supervised, respectively.

In order to compare a supervised approach and different levels of semi-supervision, we created several subsets of labeled and unlabeled instances within the training sets. This enables to analyze different levels of semi-supervision. While the size of the training set remains fixed, smaller subsets of labeled instances in it yield a rather semi-supervised approach. The size of the labeled subset ranges from 50 instances to the whole training set. Figure 3.1 shows how we split training sets into labeled and unlabeled subsets. Supervised approaches learn from the labeled subset, and ignore the unlabeled one, whereas semi-supervised approaches make predictions on the latter to increase the learning base. For each training set size, we perform 6 different selections of labeled subsets. We show the average accuracy of all 6 runs on the results.



**Figure 3.1:** Example of splitting a training set into labeled and unlabeled subsets. The former remains fixed, whereas the size of the latter two changes.

### 3.2.1 Native Multiclass Approaches

Native multiclass approaches consider the classification as a single task performed by only one classifier. They can be implemented either in a supervised or a semi-supervised basis. However, little work has been done on developing native semi-supervised approaches. The only algorithm was presented by Yajima and

Kuo (2006), which as far as we know has not been used in later works. The algorithm is not available for the community, and its implementation does not seem feasible because its description does not provide enough details to reproduce it. Thus, we propose the implementation of a semi-supervised method by following an approach similar to those by Qi et al. (2004) and Xu and Schuurmans (2005). They perform a two-step classification task, by extending the training set using a clustering algorithm in the first step. Afterward, they run a supervised SVM on the extended training set. Our approach differs from those two in that we use the same algorithm in both steps. With an approach we call *2-steps-mSVM*, we extend the training set using a supervised SVM, i.e., learning from the labeled subset, and labeling the unlabeled subset relying on classifier's decisions. We run the same algorithm on the extended set after that. As a supervised method, we use a native multiclass SVM, which we call *1-step-mSVM*.

### 3.2.2 One-Against-All Approaches

*One-against-all* is a method to split the multiclass task into smaller binary problems. Specifically, it creates  $k$  classifiers defining that many hyperplanes; each of them separates the class  $i$  from the remainder  $k-1$ . Thus, the number of classifiers is the same as the number of classes. In the test phase, all the classifiers will provide a margin for each instance, defining whether it belongs to the positive class (class  $i$ ) or the negative class (the remainder  $k-1$ ). Putting together the outputs of all classifiers for an instance (i.e., margins provided by classifiers), the one with the largest positive value will be selected as the system's decision (see Equation 3.1).

$$\hat{C}_i = \arg \max_{i=1,\dots,k} (w_i x + b_i) \quad (3.1)$$

We implemented this approach with a supervised binary SVM (*one-against-all-SVM*) and a semi-supervised binary SVM (*one-against-all-S<sup>3</sup>VM*).

### 3.2.3 One-Against-One Approaches

*One-against-one* is another method that divides a multiclass problem into smaller binary ones. Different from the above method, it creates a binary classifier for each possible pair among the  $k$  categories, what produces  $\frac{k(k-1)}{2}$  1-vs-1 classifiers. Again, a margin for all the instances is given by all the classifiers in the test phase, but the way of amalgamating the outputs changes in this case. Considering as a positive vote each time that a class beats the other in the binary classifiers, the class with most positive votes will be predicted by the system.

This method has two major problems, though:

1. As the number of classes increases, the amount of classifiers will increase in an exponential way, and so the problem could become very expensive for large taxonomies.
2. During the test phase, a 1-vs-1 classifier is unable to ignore those instances that actually belong to none of the considered pair of classes. Thus, including all the instances in the test phase is the only solution, and given that binary classifiers will provide a margin for every instance, it seems that the test phase can become noisy. This issue was also pointed out by [Chapelle et al. \(2006\)](#).

As for the one-against-all approaches, both a supervised binary SVM and a semi-supervised binary SVM were used to implement two different settings of this approach: *one-against-one-SVM* and *one-against-one-S<sup>3</sup>VM*.

### 3.3 Experiment Settings

This section introduces the datasets we have used to compare the different SVM approaches, as well as other settings.

#### 3.3.1 Datasets

In order to perform the experimentation on a multiclass scenario, we looked for suitable datasets. As benchmark datasets that have been used several times for research on classification, we chose the following:

- *BankSearch* ([Sinka and Corne, 2002](#)), a collection of 11,000 web pages over 11 classes, with very different topics: commercial banks, building societies, insurance agencies, java, c, visual basic, astronomy, biology, soccer, motor-sports and sports. We removed the category sports, since it includes both soccer and motorsports in it, and it is not at the same level as the rest of categories. This results in 10,000 web pages over 10 categories. 4,000 instances were assigned to the training set, while the other 6,000 were left on the test set.
- *WebKB*<sup>1</sup>, with a total of 4,518 documents from 4 universities, and classified into 7 classes (student, faculty, personal, department, course, project and other). The class named *other* was removed due to its ambiguity, and so we finally got 6 classes. 2,000 instances fell into the training set, and 2,518 into the test set.

---

<sup>1</sup><http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/>

- *Yahoo! Science* (Tan et al., 2002), with 788 scientific documents classified into 6 classes (agriculture, biology, earth science, math, chemistry and others). We selected 200 documents for the training set, and 588 for the test set.

Even though these cannot quite be considered as large-scale datasets, the fact that the selected training sets are small as compared to whole collections makes the problem more similar. The selection of number of instances on the training sets above depend on the number of classes and the size of each dataset.

### 3.3.2 Document Representation

SVM relies on a Vector Space Model (VSM) and thereby it requires a vectorial representation of the documents as an input for the classifier, for both train and test phases. To obtain this vectorial representation, we use the textual content of the web pages. To this end, we first converted the original HTML codes into plain text strings, removing all the HTML tags. After that, we removed a set of useless tokens, such as URLs, email addresses and stopwords from a public list<sup>2</sup>. The vectors representing the documents are composed of the remaining terms, where each dimension corresponds to a term. The weights of these terms in the vectors are defined by the TF-IDF (Term Frequency - Inverse Document Frequency) term weighting function (Salton and Buckley, 1988). In order to relax the computational cost of the task, we then removed the least-frequent terms by its document frequency; terms appearing in fewer than 0.5% of the documents were removed for the representation<sup>3</sup>. This process yielded term vectors with 8285 dimensions for *BankSearch* dataset, 3115 for *WebKB* and 8437 for *Yahoo! Science*.

### 3.3.3 Algorithmic Implementation

The 6 SVM approaches presented above require 3 different classifiers to construct them: a supervised multiclass one, and two binaries, one supervised and one semi-supervised. Taking into account that some SVM implementations are freely available for research, we looked for experimented and tested software. Among the studied alternatives, we opted to use *svm-light*<sup>4</sup> and its variants, by Thorsten Joachims (Joachims, 1998). We used supervised *svm-light* for *one-against-one-SVM* and *one-against-all-SVM* approaches, whereas *one-against-one-S<sup>3</sup>VM* and

<sup>2</sup><http://www.textfixer.com/resources/common-english-words.txt>

<sup>3</sup>0.5% was a reasonable value for the number of resources we were dealing with. However, this reduction applies to all the algorithms we compare in this work, and keeping the same reduction for all of them makes their results equally comparable while reducing the computational cost.

<sup>4</sup><http://svmlight.joachims.org>

*one-against-all-S<sup>3</sup>VM* were implemented by using semi-supervised *svm-light*. Finally, we used *svm-multiclass*<sup>5</sup> to implement *1-step-mSVM* and *2-steps-mSVM* approaches.

### 3.3.4 Evaluation Measures

Most of the results, not only in this chapter but also in the following chapters of this thesis, have to do with classification experiments. In order to evaluate their performance along this work, we use the accuracy as an evaluation measure. It has been widely used for text classification tasks, especially when it comes to multiclass problems. The value computed as the accuracy gives the percentage of correct predictions within the whole test set. We consider all the classes in the taxonomies to be equally relevant to the final performance, so that we do not consider any weightings in the evaluation process. Accordingly, a correct guess adds the same positive value on the accuracy, regardless of the class it belongs to.

Tables presenting accuracy values in this thesis show different training set sizes on each column, and different approaches or representation methods on each row. These accuracy values are emphasized in bold for outscoring performances within each table.

## 3.4 Results

This section presents the results of the experiments comparing the SVM approaches. We show the results organized by dataset, and analyze the different approaches, studying the appropriateness of multiclass or binary classifiers, as well as a supervised or a semi-supervised learning. Table 3.1 on the following page, Table 3.2 on the next page and Table 3.3 on the following page show those results for BankSearch, WebKB and Yahoo! Science datasets respectively.

### 3.4.1 Native Multiclass vs Combining Binary Classifiers

Our experiments compare two native multiclass approaches to four methods combining binary classifiers. First of all, the results clearly show that those relying on the *one-against-one* setting (i.e., *one-against-one-SVM* and *one-against-one-S<sup>3</sup>VM*) perform much worse than the rest for all the datasets. This outperformance confirms the issue we pointed out above in Section 3.2.3 on page 50, i.e., the inability of discriminating the instances that do not belong to the considered pair of classes adds noise into the decisions.

Among the other two settings, the native multiclass SVMs and the *one-against-all*, there is a clear outperformance for the former. The performance gap between

<sup>5</sup>[http://svmlight.joachims.org/svm\\_multiclass.html](http://svmlight.joachims.org/svm_multiclass.html)

BankSearch								
# of labeled instances	50	100	200	500	1000	2000	3000	4000
1-step-mSVM	.579	.706	.792	.869	.897	<b>.919</b>	<b>.925</b>	<b>.930</b>
2-steps-mSVM	<b>.628</b>	<b>.753</b>	<b>.826</b>	<b>.879</b>	<b>.898</b>	.916	.923	<b>.930</b>
one-against-all-SVM	.372	.485	.575	.697	.759	.816	.843	.855
one-against-all-S <sup>3</sup> VM	.506	.566	.621	.709	.763	.814	.842	.855
one-against-one-SVM	.311	.443	.549	.679	.744	.803	.826	.840
one-against-one-S <sup>3</sup> VM	.443	.513	.567	.668	.724	.782	.811	.840

Table 3.1: Accuracy results for the BankSearch dataset.

WebKB						
# of labeled instances	50	100	200	500	1000	2000
1-step-mSVM	<b>.600</b>	<b>.677</b>	<b>.739</b>	<b>.787</b>	<b>.810</b>	<b>.822</b>
2-steps-mSVM	.582	.667	.715	.750	.778	<b>.822</b>
one-against-all-SVM	.513	.587	.673	.744	.776	.783
one-against-all-S <sup>3</sup> VM	.592	.642	.691	.740	.773	.783
one-against-one-SVM	.488	.554	.648	.736	.775	.791
one-against-one-S <sup>3</sup> VM	.494	.579	.651	.718	.754	.791

Table 3.2: Accuracy results for the WebKB dataset.

Yahoo! Science			
# of labeled instances	50	100	200
1-step-mSVM	.682	.825	<b>.908</b>
2-steps-mSVM	<b>.687</b>	<b>.836</b>	<b>.908</b>
one-against-all-SVM	.506	.536	.630
one-against-all-S <sup>3</sup> VM	.570	.565	.630
one-against-one-SVM	.436	.483	.586
one-against-one-S <sup>3</sup> VM	.467	.514	.586

Table 3.3: Accuracy results for the Yahoo! Science dataset.

those two approaches differs depending on the dataset. Even though it is much smaller for the *WebKB* dataset than for the other two, it is undoubtedly clear that the native multiclass approach seems a better setting to face this kind of tasks. Moreover, regardless of the size of the labeled subset, the multiclass settings always outperform the others.

### 3.4.2 Supervised vs Semi-Supervised Learning

Besides these three settings, we have also compared a supervised and a semi-supervised learning for each of them in our experiments. When comparing the two analogous approaches for each setting, it can be seen that the semi-supervised ones (*2-steps-mSVM*, *one-against-all-S<sup>3</sup>VM* and *one-against-one-S<sup>3</sup>VM*) perform better than the supervised ones (*1-step-mSVM*, *one-against-all-SVM* and *one-against-one-SVM*) in most cases when it comes to the smallest labeled subsets. However, the contrary happens for larger labeled subsets, where the supervised approaches perform better. Looking at these results, it seems that the success of semi-supervised learning for multiclass classification is limited to very small labeled sets, where more instances are required in order to get a sufficient base to learn from.

Going in more depth in the native multiclass approaches, which perform the best, a similar conclusion can be drawn, especially for the largest dataset, *BankSearch*. Even though the semi-supervised *2-steps-mSVM* performs better than the supervised *1-step-mSVM* for the smallest labeled subsets, there is a slight outperformance for the latter when the labeled subset increases. In the case of *WebKB*, *1-step-mSVM* is always the best, probably because it is harder to predict correctly the unlabeled instance in the semi-supervised scenario when the taxonomy is made by closely related categories, and it adds noise in the learning phase. Finally, for *Yahoo! Science*, *2-steps-mSVM* performs slightly better, but since this dataset is quite small, it does not let us see whether *1-step-mSVM* would outperform for larger labeled subsets.

## 3.5 Discussion

In this study, we have compared the required approaches to help us determine (a) if we should use a native multiclass classifier or combine binary classifiers, and (b) whether or not including the predictions on unlabeled instances improves the performance of the classifier. This is not an exhaustive comparison study between SVM approaches for large-scale classification on multiclass taxonomies. An example of this is that we did not consider any native multiclass and semi-supervised approaches like that by [Yajima and Kuo \(2006\)](#), which we did not have access to –reasonwhy it has not been used subsequently. We have compared a set

of approaches available for research purposes instead.

### 3.6 Conclusion

In this chapter, we have analyzed a set of approaches to face a large-scale topical classification task, considering that it fulfills the conditions that (a) it is multiclass with more than two classes in the taxonomy, and (b) the labeled subset tends to be tiny as compared to the whole set to classify. Looking at these two aspects, we have compared 6 different SVM approaches, including (a) semi-supervised and supervised learning, and (b) 3 different settings, a native multiclass and 2 binary settings, *one-against-one* and *one-against-all*. With experiments over 3 different datasets, we have performed a comparison study between the different SVM approaches.

Parts of the research in this chapter have been published in [Zubiaga et al. \(2009b\)](#) and [Zubiaga et al. \(2009a\)](#).

We have also answered the following research questions in this chapter:

#### Research Question 1

*What kind of SVM classifiers should be used to perform this kind of classification tasks: a native multiclass classifier, or a combination of binary classifiers?*

We have shown the clear superiority of the native multiclass SVM classifiers over the other approaches combining binary classifiers. Our results show that relying on a set of binary classifiers is not a good option when it comes to multiclass taxonomies. Accordingly, native multiclass classifiers, which consider all the classes at the same time and have more knowledge of the whole task, perform much better.

#### Research Question 2

*What kind of learning method performs better for this kind of classification tasks: a supervised one, or a semi-supervised one?*

Semi-supervised approaches may perform better when the labeled subset is really small, but supervised approaches, which are computationally less expensive, perform similarly with more labeled documents. Therefore, we have also shown that, unlike binary tasks as shown by [Joachims \(1999\)](#), a supervised approach performs very similar to a semi-supervised approach on these environments. It seems reasonable that predicting the class of uncategorized documents is much more difficult when the number of classes increases, and so the miscategorized documents are harmful for classifier's learning.

Thereby, according to these conclusions, we decided to use a supervised multiclass SVM in this thesis, i.e., *svm-multiclass* by [Joachims \(1998\)](#). We use the



*1-step-mSVM* approach in [Chapter 5 on page 75](#), [Chapter 6 on page 95](#) and [Chapter 7 on page 111](#).



## Generation of Social Tagging Datasets

*“As a general principle, the more users share about themselves, the more others in the community will learn about them and identify with them.”*

— Matt Rhodes

This chapter describes and analyzes in detail the social tagging datasets we have created to use throughout this work. After looking for existing datasets, we found no one that fulfilled our requirements. Hence, we introduce the process we followed for generating suitable datasets, and we analyze their main characteristics.

The chapter is organized as follows. First, in [Section 4.1 on the next page](#) we describe the requirements and criteria that led us to the selection of the appropriate social tagging systems. In [Section 4.2 on page 61](#) we comprehensively analyze the features of the selected social tagging systems. Next, we present the process we carried out for gathering the datasets from the Web in [Section 4.3 on page 63](#). Then, we analyze the folksonomies of such datasets and present a set of statistics in [Section 4.4 on page 65](#). In [Section 4.5 on page 70](#) we introduce the additional data, besides tagging data, we retrieved and included in the datasets. Finally, we conclude and answer the following research question in [Section 4.6 on page 72](#):

### Research Question 3

*How do the settings of social tagging systems affect users' annotations and the resulting folksonomies?*

## 4.1 Selection of Social Tagging Systems

First of all, we defined a set of conditions that the selected social tagging systems should fulfill according to our requirements:

1. They must have a large community of users involved. This enables to further analyze the aggregation of annotations. The fact of considering whether or not a community is large can obviously be subjective, though. We consider it large enough when there is an active community and resources tend to be annotated by many users.
2. In order to gather the required data, they must provide an accessible API, or an alternative way to access the data by HTML scraping instead. The required data include full access to the triple involved in each bookmark, i.e., the user annotating it, the resource being annotated, and the tags. This is extremely relevant to analyze the nature and structure of folksonomies, and how they are created.
3. Regarding the ground truth we will assume for the classification tasks, the considered resources must somewhere be classified on consolidated taxonomies by experts. These categorization data will provide a way to quantitatively evaluate the classification tasks.

We thought it would be wise to analyze the existence of social tagging datasets that fulfilled our requirements. Even though we looked for social tagging datasets created and made publicly available by others, we just found a few of them by then, and none of them matched our needs<sup>1</sup>. Therefore, we decided to create new datasets. Before creating the datasets, though, it is of utmost importance to select the appropriate social tagging sites to collect them from. Since we wanted to analyze in depth the tagging structure of folksonomies, we were required to get data as detailed as possible. However, not all the social tagging sites provide all these data.

We analyzed a large set of social tagging sites, and studied whether or not they matched the above requirements. We found that most of them were in the long tail according to the size of the community, with small and almost inactive groups of contributors<sup>2</sup>. We ruled them out, and considered those in the head with large and active communities. Not all of them provide all the required data, though. Some social tagging sites show the aggregated list of tags for each resource, but there is no way to extract bookmark data, and thus the exhaustive

---

<sup>1</sup>By then, the only dataset with categorization data for tagged resources was CABS120k08 by Noll and Meinel (2008b), but it did not fulfill our requirements: <http://www.michael-noll.com/cabs120k08/>

<sup>2</sup>e.g., CiteULike (<http://www.citeulike.org>) is a bookmarking site for publications where usually there is not enough aggregation of annotations on a resource.

list of users who contributed and tags assigned by them when saving the resource. Moreover, some social tagging sites only show a list of tags for each resource, without the number of users annotating them<sup>3</sup>. Also, there are sites where the annotated resources have no consolidated category data<sup>4</sup>. Hence, even though there are lots of social tagging sites available online, most of them restrict the access to data, or do not fulfill all the requirements. Thereby we finally got a smaller list of social tagging sites, since lots of them had to be discarded: (i) Delicious<sup>5</sup>, where users save and annotate web pages, (ii) LibraryThing<sup>6</sup>, a social tagging site for books, and (iii) GoodReads<sup>7</sup>, also for books. In fact, all of them consist of bookmarks of resources which are regularly classified by experts. Web pages have been organized into web directories since 1990s, and librarians have been cataloging books into categories for centuries.

## 4.2 Characteristics of the Selected Social Tagging Systems

Even though all the tagging systems have the same end of enabling users to bookmark and annotate the resources of their interest, there are several features that make each of them different from the rest. The design of the interface, constraints on the inputted tags, and other features could influence users' annotations. Thus, it is worthwhile studying the nature of each of the social tagging sites we rely on, in order to understand their underlying folksonomies.

**Delicious** is a social bookmarking site that allows users to save and tag their favorite web pages, in order to ease the subsequent navigation and retrieval on large collections of annotated bookmarks. Being a social bookmarking site, every web page can be saved, so that the range of covered topics can become as wide as the Web is. It is known that the site is biased to some computer and design related topics though. Tagging web pages is one of the main features of the site, and that is the first thing the system asks for when a user saves a URL as a bookmark. The system suggests tags used earlier for that URL if some users had annotated it before. Thus, new annotators can easily select tags used by earlier users without typing them. This could encourage users to reuse others' tags, reducing the number of new tags assigned to a resource.

---

<sup>3</sup>e.g., GiveALink (<http://www.givealink.org>) only shows an unweighted list of popular tags for each bookmarked web page.

<sup>4</sup>e.g., Last.fm (<http://www.last.fm>) provides large amounts of annotations for musical groups, but there is no standard taxonomy organizing them by musical genres.

<sup>5</sup><http://delicious.com>

<sup>6</sup><http://www.librarything.com>

<sup>7</sup><http://www.goodreads.com>

**LibraryThing** and **GoodReads** are social cataloging<sup>8</sup> sites where users save and annotate books. Commonly, users annotate the books they own, they have read, or they are planning to read. We believe that users contributing to this kind of sites are more knowledgeable of the resources than those contributing to social bookmarking systems. Moreover, there are also writers and libraries contributing as users, who have a deep background on the field. This could yield annotations providing further and more detailed knowledge. The main difference among these two systems is that LibraryThing does not suggest tags when saving a book, whereas GoodReads lets the user select from tags within their personomy, that is, tags they previously assigned to other books. The latter makes it easier to reuse users' favorite tags, without re-typing them. This could encourage users to keep a smaller tag vocabulary, where they barely use new tags they did not used previously. Moreover, LibraryThing brings the user to a new page when saving a book, where they can attach tags to it; GoodReads, though, requires the user to click again on the saved book to open the form to add tags. Another remarkable difference is that LibraryThing allows some users to group tags with the same meaning, linking thus typos, misspellings, synonyms and translations to a single tag, e.g., *science-fiction*, *sf* and *ciencia ficción* are grouped into *science fiction*.

Despite of the aforementioned differences, all of them have some characteristics in common: users save resources as bookmarks, a bookmark can be annotated by a variable number of tags ranging from zero to unlimited, and the vocabulary of the tags is open and unrestricted. Table 4.1 summarizes the main features of the three social tagging sites we study in this thesis.

	<b>Delicious</b>	<b>LibraryThing</b>	<b>GoodReads</b>
<b>Resources</b>	web documents	books	books
<b>Tag suggestions</b>	based on earlier bookmarks on the resource	no	based on user's personomy
<b>Users</b>	general	readers, writers & libraries	readers, writers & libraries
<b>Tag grouping</b>	no	selected users suggest merging tags	no
<b>Vocabulary</b>	open	open	open
<b>Tag insertion</b>	space-separated	comma-separated	one by one text-box
<b>When saving a resource</b>	prompts user to add tags	prompts user to add tags at second step	user needs to click again to add tags

**Table 4.1:** Characteristics of the studied social tagging systems.

<sup>8</sup>Both social bookmarking and social cataloging refer to social tagging systems. The sole difference is on the resources, i.e., URLs are bookmarked, whereas books are cataloged.

### 4.3 Generation Process of Datasets

Even though the three chosen social tagging sites provide public access to the full bookmarking activity, getting large collections of data from them turns into a complicated task. All of them have an API for accessing the data, but none of the APIs provides the required data, so crawling the sites and scraping the HTML code instead seems to be the only way to achieve the goal. Moreover, each site sets its own limit on the number of requests, and lots of them must be done in order to obtain large-scale datasets. Hence, we set a crawling policy for each site, and applied it with extra care in order to not get banned while getting as much data as possible.

#### 4.3.1 Getting Popular Resources

As a starting point, we focused on getting a set of popular resources from each site. This provided an initial list of popular resources which represented a good seed to start the gathering process from. Those resources were also more likely to have been categorized by experts rather than resources in the tail with fewer annotations. We could also start the process by looking for popular tags or active users, but starting from resources sounds reasonable when those are what we aim to classify. Next, we will focus on the process of gathering the data in such a way that those resources are well represented insofar as involved users and their annotations are taken into account. Apart from representing those resources, we were also interested in gathering additional data, in order to represent involved users and tags to a great extent.

First of all thus we queried the three sites for popular resources. We consider a resource to be popular if at least 100 users have bookmarked it<sup>9</sup>. In the case of Delicious, we found a set of 87,096 unique URLs fulfilling this requirement. As regards to LibraryThing and GoodReads, we found an intersection of 65,929 popular books. Since the latter two rely on the same resources, we created parallel datasets for them, where the same books have categorization data attached.

#### 4.3.2 Looking for Classification Data

In the next step, we looked for classification labels assigned by experts for both kinds of resources. For the URLs gathered from Delicious, we used the Open Directory Project<sup>10</sup> (ODP) as a classification scheme. ODP is an open web directory, constructed and maintained by a community of volunteer editors, and it includes

<sup>9</sup>It was shown that the tag set of a resource tends to converge when 100 users contribute to it (Golder and Huberman, 2006). Thereby we consider it as a threshold for a resource to be popular.

<sup>10</sup><http://www.dmoz.org>

categorization data on a hierarchical structure for more than 4 million URLs. A matching between popular URLs on Delicious and those in the ODP returned a set of 12,616 URLs with a category assigned. For the set of books, we fetched their classification for both the Dewey Decimal Classification (DDC) and the Library of Congress Classification (LCC) systems. The former is a classical taxonomy that is still widely used in libraries, whereas the latter is used by most research and academic libraries. We found that 27,299 books were categorized on DDC, and 24,861 books had an LCC category assigned. In total, there are 38,148 books with category data from either one or both category schemes.

In this thesis, we will focus on both the top level and the second level of the taxonomies. This enables to evaluate the usefulness of social tags for classification on both broader and narrower categories. Even though taxonomies are made up by more than 2 levels of categorization, going into deeper levels would lack of enough number of resources for each category, and would not enable an appropriate experimentation. Table 4.2 summarizes the number of classes in each taxonomy and level, as well as the number of resources with categorization data for each of them. We kept the structure of all the taxonomies as they were, but made a little change for LCC: we merged E (*History of America*) and F (*History of the United States and British, Dutch, French, and Latin America*) categories into a single one, as it is not clear that they are disjoint categories. Also, note that the number of resources is slightly smaller for second levels. This is because we removed second-level categories and their underlying resources when there were fewer than 5 resources in them, due to the low representativity<sup>11</sup>.

	Top level		Second level	
	Resources	Classes	Resources	Classes
ODP	12,616	17	12,286	243
DDC	27,299	10	27,040	99
LCC	24,861	20	23,565	204

**Table 4.2:** Number of resources and classes for the classification experiments.

### 4.3.3 Gathering Tagging Data

Finally, we queried (a) Delicious for gathering all the personomies involved in the set of categorized URLs, and (b) LibraryThing and GoodReads for gathering all the personomies involved in the set of categorized books. By personomy, we consider the whole list of bookmarks posted by a user, including an identifier of

<sup>11</sup>The threshold of 5 resources is arbitrary. It is reasonable from our point of view, because it increases the likelihood of having more than one learning instance for each category, and the reduction of the dataset is minimal.



the resources and the tags attached by them. All three sites present no restrictions on the bookmarks shown in personomies, so that they return all available public bookmarks for the queried users.

The process above results in a large collection of bookmarks for each dataset. Within the gathered data, we focus on the following information for each bookmark:

- **User (U):** an identifier of the user who annotated the resource.
- **Resource (R):** the resource annotated by the bookmark. It is a URL in the case of Delicious, and the ISBN identifier of a book in the case of LibraryThing and GoodReads.
- **Tags (T):** the set of tags, in case it is available, annotated by the user to the resource.

That is, the triple of  $U \times R \times T$  involved in a bookmark. In this process, we consider all the tags attached to each bookmark, except for GoodReads. In this case, a tag is automatically attached to each bookmark depending on the reading state of the book: `read`, `currently-reading` or `to-read`. We do not consider this to be part of the tagging process, but just an automated step that does not provide useful information for classification, and we removed all their appearances in our dataset.

## 4.4 Statistics and Analysis of the Datasets

In order to understand the nature and characteristics of each dataset, and to analyze how the settings of each social tagging system affect the folksonomies, we study and present statistics of the created datasets.

It is worthwhile noting that, as we stated above, attaching tags to a bookmark is an optional step, so that depending on the social tagging site, a number of bookmarks may remain without tags. Table 4.3 on the next page presents the number of users, bookmarks and resources we gathered for each of the datasets, as well as the percent with attached annotations. In this work, as we rely on tagging data, we only consider annotated data, ruling out bookmarks without tags. Thus, from now on, all the results and statistics presented are based on annotated bookmarks. From these statistics, it stands out that most users (above 87%) provide tags for bookmarks on Delicious, whereas there are fewer users who tend to assign tags to resources on LibraryThing and GoodReads (roughly 38% and 17%, respectively). This shows the importance of Delicious' encouragement to adding tags, and GoodReads' discouragement to this end, requiring the user to click twice on the book in order to add tags. The latter makes the tagging

process cumbersome, and yields a large number of untagged bookmarks. LibraryThing is halfway between those two, which automatically conveys the user to the tagging form, but at a skippable second step after saving the book.

Delicious			
	Annotated	Total	Ratio
Users	1,618,635	1,855,792	87.22%
Bookmarks	273,478,137	300,571,231	91.00%
Resources	92,432,071	102,828,761	89.89%
Tags		11,541,977	-
LibraryThing			
	Annotated	Total	Ratio
Users	153,606	400,336	38.37%
Bookmarks	22,343,427	44,612,784	50.08%
Resources	3,776,320	5,002,790	75.48%
Tags		2,140,734	-
GoodReads			
	Annotated	Total	Ratio
Users	110,344	649,689	16.98%
Bookmarks	9,323,539	47,302,861	19.71%
Resources	1,101,067	1,890,443	58.24%
Tags		179,429	-

**Table 4.3:** Statistics on availability of tags in users, bookmarks, and resources for the three datasets.

The crawling process enabled us to gather large amounts of bookmarks. Not all of them correspond to the resources with categorization data from experts, though. When gathering personomies, we also gathered lots of bookmarks for resources without categorization data. Table 4.4 on the facing page shows the statistics on resources' and bookmarks' belonging to the categorized or uncategorized subset of resources, according to the categorization data we gathered from expert-driven taxonomies. It can be seen that the number of categorized bookmarks or resources is always much lower than the number of uncategorized ones. This enables to analyze a larger folksonomy as a whole for finding out tagging patterns on each site, in order to experiment afterward on the categorized subset.

A first glance at the vocabulary employed in each folksonomy can be performed by looking at the top tags on each site. The top 10 of tags set by users for each of the datasets is listed in Table 4.5 on the next page. On one hand, top tags on Delicious include tags like `design`, `software` and `blog`, showing its computer and design related bias. On the other hand, top tags on LibraryThing and GoodReads share some similarities, where tags related to literary genres stand out. Moreover, the latter shows that `non-fiction` and `nonfiction` are two of the most popular tags, whereas they appear grouped for the former.

Resources						
	Top level			Second level		
	Categ.	Uncateg.	Ratio	Categ.	Uncateg.	Ratio
Delicious (ODP)	12,616	92,419,455	0.014%	12,286	92,419,785	0.013%
LibraryThing (DDC)	23,617	3,752,703	0.629%	22,409	3,753,911	0.597%
LibraryThing (LCC)	24,861	3,751,459	0.636%	23,566	3,752,754	0.628%
GoodReads (DDC)	23,617	1,077,450	2.192%	22,409	1,078,658	2.077%
GoodReads (LCC)	24,861	1,076,206	2.310%	23,566	1,077,501	2.187%
Bookmarks						
	Top level			Second level		
	Categ.	Uncateg.	Ratio	Categ.	Uncateg.	Ratio
Delicious (ODP)	10,984,426	262,493,711	4.185%	10,773,505	262,704,632	4.101%
LibraryThing (DDC)	4,266,445	18,076,982	23.602%	4,238,774	18,104,653	23.413%
LibraryThing (LCC)	3,777,353	18,566,074	20.345%	3,607,935	18,735,492	19.257%
GoodReads (DDC)	1,615,235	7,708,304	20.954%	1,611,833	7,711,706	20.901%
GoodReads (LCC)	1,465,740	7,857,799	18.653%	1,432,073	7,891,466	18.147%

**Table 4.4:** Ratio of resources and bookmarks belonging to categorized or uncategorized data. The ratio value represents the percent of categorized bookmarks as compared to the uncategorized ones.

Regarding the distribution of tags across all the resources, users and bookmarks in the datasets, there is a clear difference of behavior among the three collections. Figure 4.1 on page 69 shows, on a logarithmic scale, the percent of resources, users and bookmarks on which tags are annotated according to their rank on the system. That is, the X axis refers to the percent of the tag rank, whereas the Y axis represents the percent of appearances in resources, users and bookmarks. For instance, if the tag ranked first had been annotated on the half of the resources, the value for the top ranked tag on resources would be 50%. Thus, these graphs enable to analyze how popular are the tags in the top as compared

Delicious	LibraryThing	GoodReads
design	fiction	fiction
blog	non-fiction	fantasy
tools	fantasy	non-fiction
software	history	own
webdesign	mystery	young-adult
web	science fiction	classics
reference	read	mystery
programming	biography	romance
music	poetry	wishlist
web2.0	novel	nonfiction

**Table 4.5:** Top 10 most popular tags on the datasets.

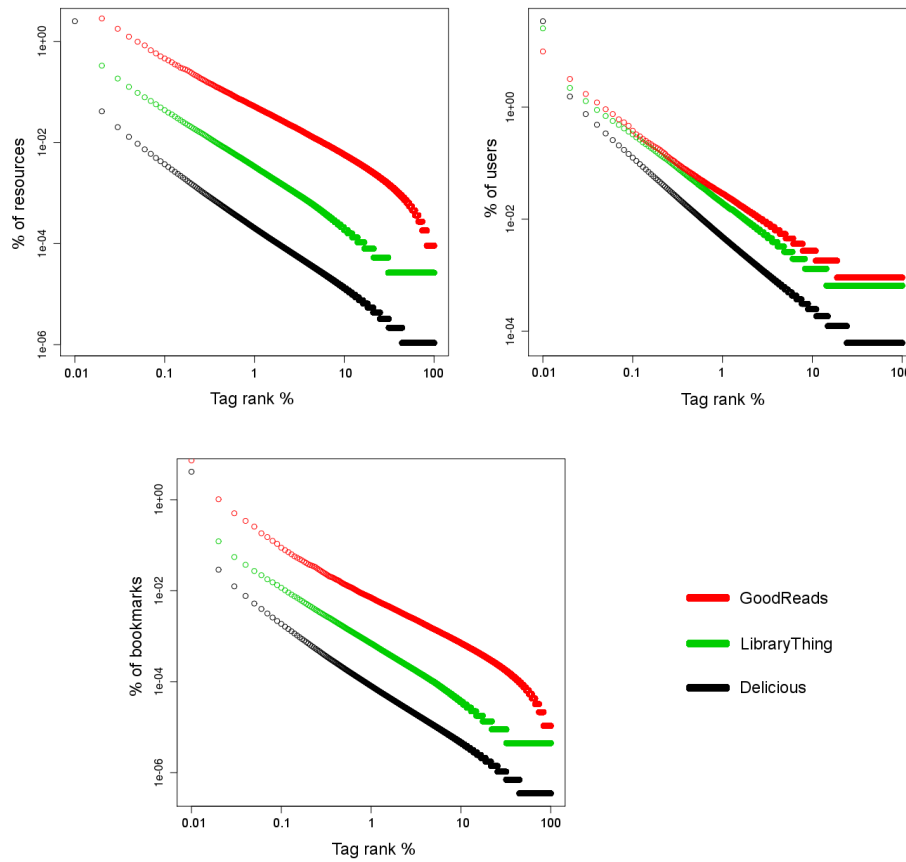
to the tags in the tail on each site. Figure 4.2 on page 70 shows the average usage of tags in a given rank for resources for each dataset. That is, we give a value of 1 to the tag annotated the most on a resource, hence ranked first for that resource. The second tag is given the value according to the fraction of users annotating it as compared to the first one. And so on for tags ranked third, fourth,... on resources. Finally, we compute the average of tags ranked on each position, which is shown in the graph. It helps infer the popularity gap between top tags on resources and tags ranked lower. Looking at those two figures, and combining their meanings, it stands out that GoodReads has the highest usage of tags in the tail, but Delicious presents the highest usage of tags in the top. Delicious is the site with highest diversity of tags, where a few tags become really popular (both in the whole collection and on resources), and many tags are seldom-used. We believe that the reasons for these differences on tag distributions are:

- Since Delicious suggests tags that have been annotated by previous users to a resource, it is obvious that those tags on the top are likely to happen more frequently, whereas others may barely be used.
- LibraryThing and GoodReads do not suggest tags used by earlier users and, therefore, tags other than those in the top tend to be used more frequently than on Delicious.
- GoodReads suggests tags from previous bookmarks of the same user, instead of tags that others assigned to the resource being tagged. Thus, this encourages reusing tags in their personomy, making it remain with a smaller number of tags (see Table 4.6). In addition, users tend to assign fewer tags to a bookmark on average, probably due to the one-by-one tag insertion method of site's interface.

# of tags	Delicious	LibraryThing	GoodReads
Per resource	33.35	14.53	13.33
Per user	632.714	357.15	131.03
Per bookmark	3.75	2.46	1.55

**Table 4.6:** Average counts of different tags.

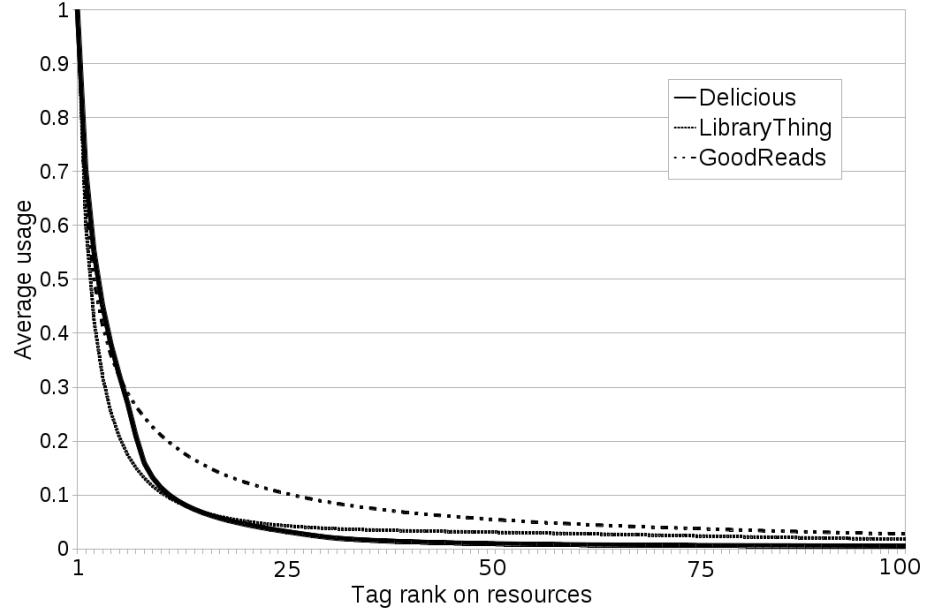
Regarding the distribution of tags across resources, users, and bookmarks, Figure 4.3 on page 71 shows percents of tags appearing more, equal or less frequently in an item (i.e., resources, users or bookmarks) than in another. It is obvious that a tag cannot appear in a smaller number of bookmarks than users or resources, by definition. Looking at the rest of data, it stands out that tags tend to appear in more bookmarks than users ( $b > u$ ) and more resources than users ( $r > u$ ) for GoodReads, due to the same feature that allows users to select among



**Figure 4.1:** Tag usage percentages in the collection. These 3 graphs represent, on a logarithmic scale for both  $x$  and  $y$  axes, the percent of annotations to resources, users, and bookmarks per tag rank.

tags in their personomy. However, LibraryThing and Delicious have many tags present in the same number of bookmarks and users ( $b = u$ ), and resources and users ( $r = u$ ), even though the difference is more marked for the former site. This reflects the large number of tags that users utilize just once on these sites. All three sites have two features in common: there are a few exceptions of tags utilized by more users than the number of resources it appears in ( $r < u$ ), and almost all the tags are present in the same number of bookmarks and resources ( $b = r$ ). The latter, combined with the lower ( $b = u$ ) values, means there is a large number of users spreading personal tags across resources that only have a bookmark with that tag, especially on GoodReads, but also for the other two sites.

Finally, we analyze to what extent a bookmark introduces new tags into a resource that were not present in earlier bookmarks. Figure 4.4 on page 72 shows

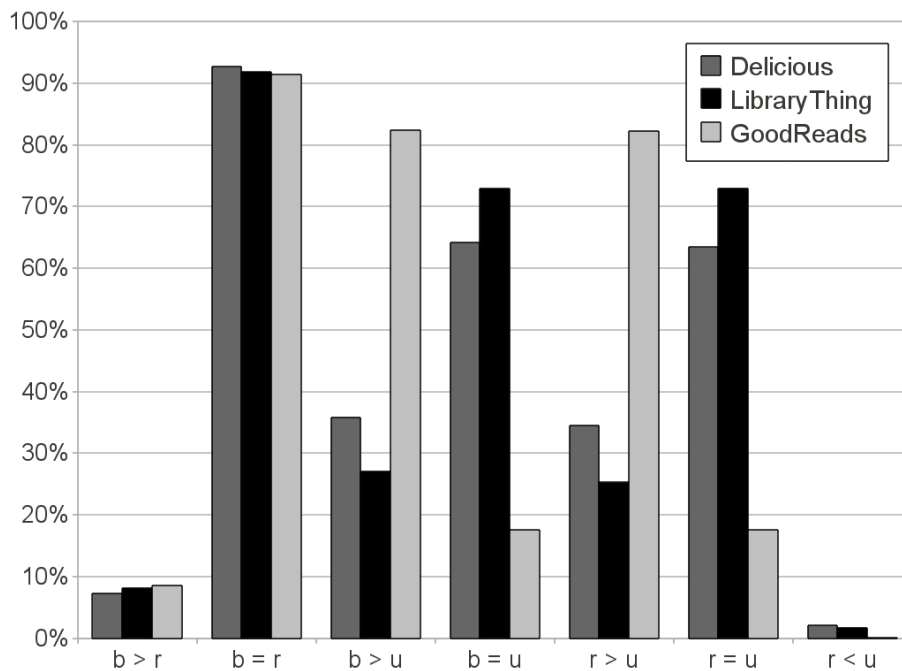


**Figure 4.2:** Tag usage percentages on resources. Each tag rank represents the average usage of tags appearing in that position on resources as compared to the top ranked tag.

these statistics for Delicious and LibraryThing. The same graph for GoodReads is not shown because neither the timestamp nor the ordering of the bookmarks is available in our dataset. The graph shows, on average, the ratio of new tags, not present in earlier bookmarks of a resource, assigned in bookmarks that rank from first to 100th bookmark, i.e., if  $tag_1$  and  $tag_2$  were annotated in the first bookmark of a resource, and  $tag_2$  and  $tag_3$  in the second bookmark for the same resource, the ratio of novelty for the second bookmark is of 50%. It stands out the marked inferiority of tag novelty on Delicious as against to LibraryThing. This is, again, due to the tag suggestion policy of Delicious, what brings about a higher likelihood of reusing previously existing tags.

## 4.5 Gathering Additional Data

Besides all the aforementioned tagging data, we also gathered some more data about the categorized resources. We needed other data sources in order to perform comparisons with tagging data along the experimentation. Specifically, we compare the usefulness of tagging data as against to other sources for the classi-



**Figure 4.3:** Tag distribution across resources (r), users (u) and bookmarks (b). Each bar represents the percent of tags that match the condition on X axis.

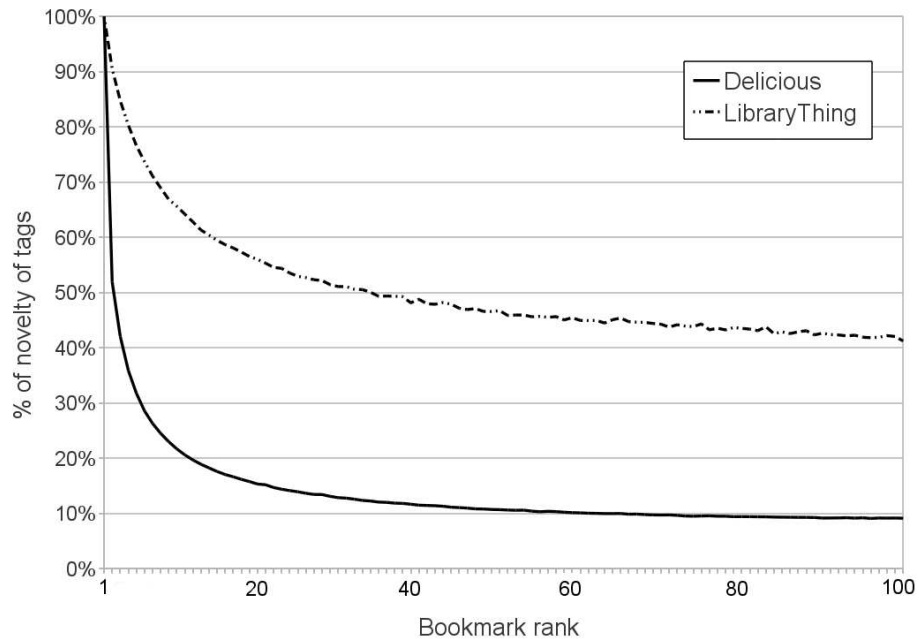
fiction task in Chapter 5 on page 75, and also require additional data to analyze the descriptiveness of tags in Chapter 7 on page 111.

On one hand, we got the following data for the categorized URLs:

- **Self-content:** it is the content of the web page itself, i.e., the HTML code fetched from the original URL.
- **Notes:** a note can be defined as a free text describing the content of a web page. It is available on Delicious, and it is intended to provide a means to briefly summarize the aboutness of a web page.
- **Reviews:** a review may be considered to be fairly similar to notes. However, reviews as they were collected from StumbleUpon<sup>12</sup>, usually have a subjective bias, where users tend to value how they like the content of a web page.

On the other hand, with regard to the categorized books, there is no easy way to get the content of the book. We did not have access to the books, since most of them are not freely available. Thus, we got the following metadata associated to the books:

<sup>12</sup><http://www.stumbleupon.com>



**Figure 4.4:** Novelty ratio of tags per rank of bookmark.

- **Synopses:** a synopsis is a brief summary of the content of a book, which is usually printed on the back cover. We fetched synopses from the book retailer Barnes&Noble<sup>13</sup>.
- **Editorial reviews:** summaries written by the publisher, or other professionals, are considered as editorial reviews. We gathered them from Amazon<sup>14</sup>.
- **User reviews:** we also collected reviews written by users on LibraryThing, GoodReads and Amazon, where they comment on the books with their summaries and thoughts.

Since we do not have access to the self-content of the books, we will consider both synopses and editorial reviews as a summary of their contents.

## 4.6 Conclusion

We have studied the characteristics of several social tagging systems, and concluded with three sites that fulfill our requirements: Delicious, LibraryThing and GoodReads. We have created three large-scale social tagging datasets from these sites including millions of bookmarks, not only for web pages, but also for books,

<sup>13</sup><http://www.barnesandnoble.com>

<sup>14</sup><http://www.amazon.com>



which enables further analyzing other kinds of resources. To the best of our knowledge, these are the largest social tagging datasets used for research so far. Also, we have analyzed the statistics of the datasets and the features of the underlying folksonomies.

Even after we created these social tagging datasets, and made publicly available parts of them<sup>15</sup>, little work has been done on creating more datasets and especially on releasing them. In Körner and Strohmaier (2010), the authors present a list of publicly available social tagging datasets, among which our datasets are also included. However, the authors set out the problem of the unavailability of more datasets, and encourage researchers to create and release new ones.

In this chapter, we have answered the following research question:

### Research Question 3

*How do the settings of social tagging systems affect users' annotations and the resulting folksonomies?*

To this end, we have analyzed several features that can be found in different settings of social tagging systems. Among the analyzed features, we have shown the impact of tag suggestions, which considerably alters the resulting folksonomy. In the studied social tagging sites, all of them differ on the settings regarding suggestions:

- **Resource-based suggestions (Delicious):** when the system suggests tags assigned by other users to the resource at the time of bookmarking it, the likelihood of using new tags to further describe such a resource decreases. In this case, users provide less originality and tend to rely on system suggestions.
- **Personomy-based suggestions (GoodReads):** when the system suggests tags previously used by the user, the vocabulary in their personomy tends to be much smaller. However, users do not know how others annotated a resource, and thus they are likely to provide new tags to the resource.
- **Without suggestions (LibraryThing):** when the system does not suggest any tags to the user, the vocabulary in their personomy increases, as well as the diversity of tags in each resource.

From now on, in Chapter 5 on page 75, Chapter 6 on page 95 and Chapter 7 on page 111, we will use these three datasets for experimentation, and we will analyze in more depth their features and how they affect the performance of a resource classification task.

---

<sup>15</sup><http://nlp.uned.es/social-tagging/datasets/>



## Representing the Aggregation of Tags

*“If we wish to discuss knowledge in the most highly developed contemporary society, we must answer the preliminary question of what methodological representation to apply to that society.”*

— Jean-Francois Lyotard

In this chapter, we set out to propose and evaluate different representations of resources based on social tags for a resource classification task. Each user contributing to the annotations on a resource provides their own tags, which commonly differ from others'. We explore different ways of representing large amounts of annotations provided by users and aggregated on resources on social tagging systems. We also measure the potential of social tags as compared to other data sources including self-content and user reviews, and analyze the suitability of combining them in search of a better performance of the classifier.

This chapter is organized as follows. Next, in [Section 5.1 on the following page](#) we describe the way user annotations are aggregated on a resource to go into the problem. In [Section 5.2 on page 77](#) we propose several representation approaches for social tags. Then, we present the results of the tag-based classification in [Section 5.3 on page 79](#), and compare them to the results by other data sources in [Section 5.4 on page 84](#). We describe the experiments on combining data sources in [Section 5.5 on page 86](#), and conclude the chapter in [Section 5.6 on page 90](#).

The following research questions are addressed in this chapter:

**Research Question 4**

*What is the best way of amalgamating users' aggregated annotations on a resource in order to get a single representation for a resource classification task?*

**Research Question 5**

*Despite of the usefulness of social tags for these tasks, is it worthwhile considering their combination with other data sources like the content of the resource as an approach to improve the results even more?*

**Research Question 6**

*Are social tags also useful and specific enough to classify resources into narrower categories as in deeper levels of hierarchical taxonomies?*

## 5.1 Aggregation of User Annotations

Social tagging systems allow users to annotate on resources that others have previously annotated. This enables the aggregation of annotations provided by many users on the same resource. Obviously, each user provides their own annotations, so that tags tend to be different from user to user. These annotations are listed all together in a detailed manner (see Table 5.1), and merged into a single list of top tags which summarizes the Full Tagging Activity (in the following, FTA) on a resource (see Table 5.2 on the next page).

User annotations: Flickr.com	
<b>User 1:</b>	photo, photography, images, pictures
<b>User 2:</b>	photo, web2.0, social, tools, blog
<b>User 3:</b>	cloud, pictures, sharing
<b>User 4:</b>	flickr, photos
<b>User 5:</b>	photo, sharing, tool

**Table 5.1:** Example of annotations for the URL Flickr.com on the social bookmarking site Delicious.

The tagging activity of a community of users on a resource creates an aggregated list of tags. A resource annotated by  $p$  users will have a list of  $n$  different tags, where each tag could have been assigned by  $p$  or fewer users. The number of users who used a certain tag,  $w_t$ , defines a weight that allows to infer an ordered list of tags.

Top tags: Flickr.com (79,681 users)	
photos	22,712
flickr	19,046
photography	15,968
photo	15,225
sharing	10,648
images	9,637
web2.0	9,528
community	4,571
social	3,798
pictures	3,115

**Table 5.2:** Example of top tags for the URL Flickr.com on the social bookmarking site Delicious: the number associated to each tag represents the number of users annotating it.

This aggregation of social tags was suggested as a means to feed the classification of resources (Noll and Meinel, 2008a). However, to the best of our knowledge, no research work has been conducted on their application to an automated classifier. Moreover, it is not clear what is a good way to represent the aggregation of tags. In this chapter, we will focus on these issues by proposing, analyzing and evaluating different representations for social tags so as to classifying resources, and also comparing their performance to other data sources including self-content and user reviews. We perform such a study on two different levels of the taxonomies, exploring thereby the suitability of social tags for broader and narrower categories.

## 5.2 Representing Resources Using Tags

We believe there are two major factors that should be considered for the representation of resources using social tags provided by users: (1) the selection of the tags that should be taken into account for the representation, and (2) the weights that should be assigned to those tags.

On the one hand, as regards to the selection of tags, one could think that not all the tags are useful for the representation, but just those in the top that most users have chosen. An important feature of social tagging systems is the ability of users to coincide on some tags provided by others. Thus, the coincidence of user annotations, which is reflected on the top tags, can be considered as a consensus of the main tags that better fit the description of the resource. However, the diversity on the annotations can also give users the opportunity to assign seldom-used tags that further detail the resource. The latter could encourage considering

even tags in the tail. We will thus explore both tags in the top and the whole set.

On the other hand, the weight of each tag must be defined appropriately. We propose 4 different ways of assigning those weights:

- **Tag ranks:** the weight is assigned according to the position of a tag in the ranked top of tags. Tags corresponding to the top 10 list of a resource are assigned a value in a rank-based way. The first-ranked tag is always set the value 1, 0.9 for the second, 0.8 for the third, and so on. This approach respects the position of each tag in the top 10, but the different gaps among tag weights are ignored.
- **Tag fractions:** the weight is computed according to the fraction of users who annotate a tag,  $w_t/p$ , i.e., the number of users annotating a tag on a resource, divided by the total number of users who annotated the resource. Taking into account both the number of users who bookmarked a resource  $r$  and the weight of each tag  $w_t$ , it is possible to define the fraction of users assigning each tag. A tag would have been annotated by the totality of the users when its weight matches the user count of a resource, getting a value of 1 as the fraction. According to this, the value set to each tag is higher than 0 (since the considered tags have annotated by at least one user), and can be up to 1. This representation approach is similar to that by [Noll and Meinel \(2008a\)](#) for their analysis of the similarity between social tags and the classification by experts. However, they ignore the least popular tags by giving a value of 0, what may give rise to the removal of several tags from the representation.
- **Unweighted:** in a binary way, the presence of a tag represents a value of 1, and its absence a value of 0. The only feature considered for this representation is the occurrence or non-occurrence of a tag in the annotations of a resource. This approach thereby ignores the weights of tags, and assigns a binary value to each feature in the vector.
- **Weighted according to user counts:** it considers the number of users assigning the tag ( $w_t$ ) as a weight. The weight for each of the tags of a resource ( $w_1, \dots, w_n$ ) is considered as it is in this approach. Now, by definition, the weights of the tags are fully respected, although the amount of users bookmarking a resource is ignored. Note that different orders of magnitude are mixed up now, since the count of bookmarking users range within very different values. For instance, [Ramage et al. \(2009\)](#) used this approach in their work for clustering web pages, but they assumed it without comparing it to other representations.

Table 5.3 on the facing page shows an example of annotations on a resource, and how each of the 4 weighting measures would look like for the example.

	FTA								
	Top 10								
	$t_1$	$t_2$	$t_3$	...	$t_9$	$t_{10}$	$t_{11}$	...	$t_n$
<b>Ranks</b>	1	0.9	0.8	...	0.2	0.1	0	...	0
<b>Fractions</b>	0.5	0.3	0.2	...	0.02	0.01	0.01	...	0.01
<b>Unweighted</b>	1	1	1	...	1	1	1	...	1
<b>Weighted</b>	50	30	20	...	2	1	1	...	1

**Table 5.3:** Example of the 4 representations of social tags on a resource annotated by 100 users, and tags ranked 1st, 2nd and 3rd were annotated by 50, 30 and 20 users, respectively.

Taking into account the factors above, we propose and analyze the 7 representation approaches summarized in Table 5.4. All 4 weighting measures are included, as well as two selections of tags: the FTA including the whole set of tags, and the top 10 tags of each resource including the best-weighted ones<sup>1</sup>. In the case of the rank-based weighting, we only apply it to the top 10 of tags, because it is defined to give a weight for only 10 tags.

	Top 10	FTA
<b>Tag ranks</b>	x	
<b>Tag fractions</b>	x	x
<b>Unweighted</b>	x	x
<b>Weighted</b>	x	x

**Table 5.4:** Summary of tag representations.

## 5.3 Tag-based Classification

According to the experimental results in Chapter 3 on page 47, we have used a multiclass SVM algorithm to perform the classification tasks, feeding the classifier with social tags from the three datasets introduced in Chapter 4 on page 59. We got different sizes of training sets for each dataset, and generated 6 different random selections for each size. We present the accuracy results corresponding to the average of those 6 runs. Results are split into separate tables, with a table corresponding to each dataset (Delicious, LibraryThing, GoodReads). Each table

<sup>1</sup>We selected the top 10 because it is usual to find that number of tags on social tagging systems. However, we could have chosen another value instead, yielding comparable conclusions. We provide additional results and information on this in Appendix A on page 143.

includes results for all 7 representations introduced above, and both top and second levels of the corresponding taxonomy.

Delicious - ODP							
Top level							
	600	1400	2200	3000	4000	5000	6000
Tag Ranks	.462	.501	.493	.501	.498	.501	.484
Tag Fractions (Top 10)	.430	.447	.456	.467	.466	.462	.464
Tag Fractions (FTA)	.442	.463	.457	.460	.461	.461	.461
Unweighted Tags (Top 10)	.505	.510	.512	.517	.520	.522	.531
Unweighted Tags (FTA)	.530	.556	.566	.572	.569	.571	.572
Weighted Tags (Top 10)	.509	.576	.606	.625	.638	.645	.654
Weighted Tags (FTA)	.533	.600	.629	.647	.660	.669	.680
Second level							
	600	1400	2200	3000	4000	5000	6000
Tag Ranks	.292	.332	.345	.349	.351	.349	.360
Tag Fractions (Top 10)	.262	.280	.297	.304	.315	.317	.349
Tag Fractions (FTA)	.249	.279	.294	.308	.302	.302	.336
Unweighted Tags (Top 10)	.315	.340	.354	.351	.348	.365	.361
Unweighted Tags (FTA)	.411	.480	.502	.509	.519	.509	.529
Weighted Tags (Top 10)	.342	.432	.475	.497	.517	.532	.545
Weighted Tags (FTA)	.359	.453	.498	.522	.541	.556	.568

**Table 5.5:** Accuracy results for tag-based web page classification.

Table 5.5 shows the results on the Delicious dataset. At a first glance, it is clear that rank-based and fraction-based approaches perform much worse than the rest. Among the others, the weighted approach performs better than the unweighted one, so that considering the number of users assigning each tag seems to be the best option. Accordingly, considering the total number of users annotating a resource does not seem helpful, as shown by the underperformance of the fraction-based approach.

The unweighted approach may perform better than the weighted one for small training sets when it comes to the second level classification. It seems reasonable that the weighted approach requires more training instances to correctly represent the large diversity of possible values, and especially when the number of categories increases, as it happens on the second level. This is reflected in the underperformance of the weighted approach for the smaller training sets upon the second level of the taxonomy. However, the outperformance of the weighted approach becomes clear when the size of the training set increases.

In most cases, FTA outperforms the top 10, even though the gap is not very large. This shows that top tags are the most useful, but the rest may also be helpful to a lesser extent. Accordingly, tags in the tail chosen by fewer users provide useful data that should not be discarded. The weighted approach on all



the tags performs the best for Delicious.

Table 5.6 on the following page shows the results for the LibraryThing dataset, for DDC and LCC taxonomies. Similar to Delicious, the weighted one relying on the FTA is the outperforming approach, for both the top and second levels, and for both schemes.

Different from Delicious, the FTA-based approaches are not always better than those based on the top 10. This difference happens when using the unweighted approach for the top level classification. Many LibraryThing users tend to use personal tags describing whether or not they own the book (e.g., `own`), and the physical location of it (e.g., `a1`). These tags are barely used within a book, as personal tags that spread across books by a single user. Thus, ignoring tag weights and giving all of them the same weight overrates those personal and low-ranked tags. This may increase the likelihood of books containing a certain personal tag to be mispredicted for the same category by the classifier. Accordingly, this is the main reason for the slight gap between the top 10 and FTA based representations for the weighted approach. Tags below the top 10 are not as useful as on Delicious. Fortunately, the weighted approach underrates such personal tags, and the classifier is able to discriminate them, profiting from some low-ranked tags to slightly improve the performance. This outperformance is larger on the second level, suggesting that low-ranked tags provide more detailed description, and rather help for deeper classification.

Even though the weighted approach is the best in this case, using rank-based weights performs better than the unweighted approach, different from Delicious. After all, the weighted approach performs the best also for this dataset.

Table 5.7 on page 83 shows the results for the GoodReads dataset, for DDC and LCC taxonomies. Again, the FTA-based weighted approach is the best one, with clearly outperforming results for both top and second levels on both schemes, DDC and LCC. Different from LibraryThing, though, FTA-based approaches perform better than top 10 based ones in most cases. This shows that GoodReads users tend to use fewer personal biased tags, making low-ranked tags much more useful than for LibraryThing. Despite these differences, the weighted approach is clearly the best approach for this dataset as well.

For both LibraryThing and GoodReads, the results look very similar for both taxonomies, DDC and LCC. Even though the results are slightly better for the former, both yield similar conclusions when comparing the gaps between representation approaches. This strengthens the usefulness of the weighted approach regardless of the taxonomy being considered.

Summarizing the results for the three datasets, the FTA-based weighted approach has shown to be the best. Even though not every low-ranked tag seems useful for the classification task, the weighted approach is able to establish their representativity to the resource, getting the best results by using all the tags.

LibraryThing - DDC							
Top level							
	3000	6000	9000	12000	15000	18000	21000
Tag Ranks	.791	.783	.778	.782	.788	.787	.797
Tag Fractions (Top 10)	.719	.717	.720	.721	.727	.721	.724
Tag Fractions (FTA)	.700	.696	.701	.702	.706	.701	.706
Unweighted Tags (Top 10)	.756	.763	.753	.766	.759	.759	.758
Unweighted Tags (FTA)	.624	.622	.628	.629	.629	.628	.624
Weighted Tags (Top 10)	.858	.861	.862	.865	.866	.866	.864
Weighted Tags (FTA)	.861	.864	.864	.867	.869	.869	.868
Second level							
	3000	6000	9000	12000	15000	18000	21000
Tag Ranks	.520	.520	.526	.530	.527	.525	.532
Tag Fractions (Top 10)	.511	.513	.511	.513	.513	.517	.521
Tag Fractions (FTA)	.465	.474	.469	.470	.470	.472	.477
Unweighted Tags (Top 10)	.507	.538	.538	.532	.543	.528	.539
Unweighted Tags (FTA)	.515	.533	.530	.533	.538	.536	.538
Weighted Tags (Top 10)	.679	.687	.696	.696	.701	.701	.704
Weighted Tags (FTA)	.690	.700	.707	.709	.715	.712	.715
LibraryThing - LCC							
Top level							
	3000	6000	9000	12000	15000	18000	21000
Tag Ranks	.783	.790	.788	.783	.789	.795	.790
Tag Fractions (Top 10)	.739	.740	.741	.743	.741	.738	.746
Tag Fractions (FTA)	.711	.715	.715	.717	.714	.712	.719
Unweighted Tags (Top 10)	.759	.772	.764	.771	.763	.770	.763
Unweighted Tags (FTA)	.654	.660	.661	.661	.658	.655	.661
Weighted Tags (Top 10)	.852	.854	.856	.858	.858	.855	.858
Weighted Tags (FTA)	.853	.857	.856	.861	.861	.857	.861
Second level							
	3000	6000	9000	12000	15000	18000	21000
Tag Ranks	.519	.511	.515	.518	.512	.511	.520
Tag Fractions (Top 10)	.414	.413	.413	.415	.417	.411	.417
Tag Fractions (FTA)	.408	.409	.408	.410	.410	.409	.410
Unweighted Tags (Top 10)	.542	.568	.564	.565	.579	.550	.576
Unweighted Tags (FTA)	.596	.612	.608	.616	.615	.606	.614
Weighted Tags (Top 10)	.687	.710	.716	.720	.721	.722	.727
Weighted Tags (FTA)	.703	.725	.729	.734	.734	.736	.739

Table 5.6: Accuracy results for tag-based book classification (LibraryThing).

GoodReads - DDC							
Top level							
	3000	6000	9000	12000	15000	18000	21000
Tag Ranks	.652	.656	.659	.654	.650	.655	.668
Tag Fractions (Top 10)	.660	.658	.662	.663	.671	.659	.664
Tag Fractions (FTA)	.654	.653	.657	.658	.665	.655	.659
Unweighted Tags (Top 10)	.647	.645	.643	.650	.639	.657	.647
Unweighted Tags (FTA)	.635	.638	.637	.639	.639	.642	.640
Weighted Tags (Top 10)	.728	.730	.736	.742	.739	.740	.740
Weighted Tags (FTA)	.745	.747	.754	.757	.757	.757	.756
Second level							
	3000	6000	9000	12000	15000	18000	21000
Tag Ranks	.435	.439	.434	.447	.445	.443	.447
Tag Fractions (Top 10)	.445	.450	.450	.452	.452	.453	.458
Tag Fractions (FTA)	.432	.440	.439	.440	.440	.441	.445
Unweighted Tags (Top 10)	.430	.440	.441	.443	.435	.440	.449
Unweighted Tags (FTA)	.450	.460	.447	.454	.453	.458	.452
Weighted Tags (Top 10)	.487	.500	.503	.505	.507	.508	.510
Weighted Tags (FTA)	.509	.520	.528	.528	.530	.529	.530
GoodReads - LCC							
Top level							
	3000	6000	9000	12000	15000	18000	21000
Tag Ranks	.625	.636	.629	.630	.632	.630	.631
Tag Fractions (Top 10)	.657	.664	.665	.667	.667	.663	.674
Tag Fractions (FTA)	.650	.658	.656	.658	.659	.654	.663
Unweighted Tags (Top 10)	.625	.626	.633	.633	.634	.623	.629
Unweighted Tags (FTA)	.642	.648	.653	.651	.647	.639	.653
Weighted Tags (Top 10)	.700	.711	.711	.714	.713	.713	.721
Weighted Tags (FTA)	.725	.731	.737	.738	.734	.731	.743
Second level							
	3000	6000	9000	12000	15000	18000	21000
Tag Ranks	.404	.411	.410	.403	.404	.405	.407
Tag Fractions (Top 10)	.412	.421	.426	.427	.430	.427	.427
Tag Fractions (FTA)	.418	.427	.431	.432	.433	.432	.433
Unweighted Tags (Top 10)	.414	.419	.420	.415	.414	.422	.435
Unweighted Tags (FTA)	.462	.475	.467	.478	.477	.481	.484
Weighted Tags (Top 10)	.467	.479	.487	.486	.491	.491	.493
Weighted Tags (FTA)	.494	.507	.510	.514	.513	.517	.519

Table 5.7: Accuracy results for tag-based book classification (GoodReads).

Thereby, the aggregation of user annotations has shown to be crucial to define the representativity of a tag with respect to the annotated resource.

## 5.4 Comparing Social Tags to Other Data Sources

After we got the best representation approach to perform the classification experiments using social tags, we aimed at comparing their performance to that by other data sources. As we introduced previously in Chapter 4 on page 59, we gathered additional data for the resources we are working on, i.e., web pages and books. In both cases, we tried to gather two more types of data: content and reviews. Regarding web pages, we rely on the textual content contained in the HTML source and user reviews fetched from social networks. In the case of books, we consider synopses and editorial reviews as a summary of their content, and user-generated reviews on the other hand.

With those content and user reviews, we created a representation based on the bag-of-words model (Harris, 1970). We merged all the texts available for each source, and created a single text with them. In order to clean up those texts, we stripped HTML tags, removed stop-words and stemmed the remaining words (Porter, 1980). Then, we weighted the words according to the TF-IDF scheme. The final representation of a resource, either based on content or user reviews, is a vector composed by words weighted by their TF-IDF values.

We use the same method as above for the creation of different training set sizes with 6 runs. As both LibraryThing and GoodReads work on the same books, the content and user reviews are the same in these cases, so that we group their results into a single table. For the three datasets we work with, we show the results of using content and comments, and compare them to the best tag-based approach, that is, the FTA-based weighted approach.

Delicious - ODP							
Top level							
	600	1400	2200	3000	4000	5000	6000
Content	.518	.561	.579	.588	.595	.604	.610
Reviews	.520	.578	.602	.618	.630	.639	.646
Tags	.533	.600	.629	.647	.660	.669	.680
Second level							
	600	1400	2200	3000	4000	5000	6000
Content	.337	.394	.422	.437	.450	.464	.470
Reviews	.349	.423	.459	.478	.497	.511	.524
Tags	.359	.453	.498	.522	.541	.556	.568

**Table 5.8:** Accuracy results comparing different data sources on web page classification.

Table 5.8 on the preceding page shows the results for the Delicious dataset. In this case, self-content of web pages is the worst data source out of the three we studied. Results by self-content are far below from those by reviews and tags. Likewise, social tags are clearly the best data source for the classification task. There is a clear outperformance of tags for the top level, but the difference is even larger for the second level. This strengthens one of the main motivations of this thesis, i.e., the fact that self-content is not always representative of its aboutness, and other data sources can provide more accurate definitions.

LibraryThing & GoodReads - DDC							
Top level							
	3000	6000	9000	12000	15000	18000	21000
Content	.767	.792	.802	.809	.809	.815	.817
Reviews	.777	.808	.820	.831	.833	.839	.840
Tags (LibraryThing)	.861	.864	.864	.867	.869	.869	.868
Tags (GoodReads)	.745	.747	.754	.757	.757	.757	.756
Second level							
	3000	6000	9000	12000	15000	18000	21000
Content	.572	.612	.631	.643	.649	.657	.660
Reviews	.582	.628	.651	.667	.678	.685	.693
Tags (LibraryThing)	.690	.700	.707	.709	.715	.712	.715
Tags (GoodReads)	.509	.520	.528	.528	.530	.529	.530
LibraryThing & GoodReads - LCC							
Top level							
	3000	6000	9000	12000	15000	18000	21000
Content	.767	.789	.798	.803	.806	.807	.810
Reviews	.780	.803	.816	.823	.827	.828	.833
Tags (LibraryThing)	.853	.857	.856	.861	.861	.857	.861
Tags (GoodReads)	.725	.731	.737	.738	.734	.731	.743
Second level							
	3000	6000	9000	12000	15000	18000	21000
Content	.579	.620	.645	.658	.668	.673	.681
Reviews	.581	.637	.664	.683	.698	.705	.712
Tags (LibraryThing)	.703	.725	.729	.734	.734	.736	.739
Tags (GoodReads)	.494	.507	.510	.514	.513	.517	.519

**Table 5.9:** Accuracy results comparing different data sources on book classification.

Table 5.9 shows the results for books, using tags from LibraryThing and GoodReads. In this case, we got results similar to Delicious when using tags from LibraryThing. Again, user reviews outperform the content (although we considered synopses and editorial reviews as a summary of the content of the book in this case). Moreover, social tags perform even better than user reviews, especially for the second level classification. The results are comparable for both

classification schemes, DDC and LCC.

However, using tags from GoodReads is not enough to achieve results as good as using content or user reviews. GoodReads tags clearly underperform the other data sources. For this dataset, reviews are the data source scoring the best results. We believe that this happens because most GoodReads users do not provide tags when bookmarking a book<sup>2</sup>. This way, a community providing fewer annotations gives rise to a less accurate aggregation of tags.

Summarizing, tags show to be really powerful as compared to other data sources like the content of the resource, or user reviews on it. However, large amounts of annotations are necessary in order to score outperforming results.

## 5.5 Getting the Most Out of All Data Sources

Even though the tag-based representation outperforms in most cases the other two data sources, namely content and user reviews, all of them yield encouraging results and look good enough to combine them and try to improve even more the classifier's performance. The following questions arise from this statement: what if a classifier is guessing correctly while the others are making a mistake? Could we combine the predictions to get the most out of each of them?

An interesting approach to combine SVM classifiers is known as classifier committees (Sun et al., 2004). Classifier committees rely on the predictions of several classifiers, and combine them by means of a decision function, which serves to define the weight or relevance of each classifier in the final prediction. After applying the decision function on the predictions of all classifiers, a single unified prediction can be inferred.

An SVM classifier outputs a margin for each resource over each class in the taxonomy, meaning the reliability to belong to that class. The class with the largest positive margin for each resource is then selected as the classifier's prediction. The larger is the gap between the largest positive and the rest of margins, the more reliable can be considered the classifier's prediction. Thus, combining the predictions of SVM classifiers could be done by means of adding up their margins or reliability values for each class. Each resource will then have a new reliability value for each class, i.e., the sum of margins by different classifiers for a resource. Nonetheless, in this case, since each of the three classifiers work with different type of data, the range and scale of the margins they output differ. To solve this, we propose the normalization of the margins based on the maximum

---

<sup>2</sup>GoodReads does not encourage users to add tags as LibraryThing does, requiring a second click from the user, what brings about a large set of unannotated bookmarks, representing a ratio of more than 80% bookmarks (see Section 4.4 on page 65)

margin value outputted by each classifier,  $\max(m_i)$  (see Equation 5.1).

$$m'_{ijc} = \frac{m_{ijc}}{\max(m_i)} \quad (5.1)$$

where  $m_{ijc}$  is the margin by the classifier  $i$  between the resource  $j$  and the hyperplane for the class  $c$ , and  $m'_{ijc}$  is its value after normalizing it.

The class maximizing this sum of margins will be predicted by the classifier. Then, the sum of margins between the class  $c$  and the resource  $j$  using a committee with  $n$  classifiers is defined by Equation 5.2.

$$S_{jc} = \sum_i^n m_{ijc} \quad (5.2)$$

If the classifiers are working over  $k$  classes, then the predicted class for the resource  $j$  will be defined by Equation 5.3.

$$C_j^* = \arg \max_{i=1..k} (S_{ji}) \quad (5.3)$$

As a toy example of the possible advantage of using classifier committees, Table 5.10 on the following page shows the outputs in the form of margins of two classifiers for a resource in a taxonomy with 3 categories. Let this resource belong to the category #2. The example shows that, on one hand, the classifier A has predicted the category #1, with a margin of 1.2, but a slight gap to the category #2 which gets a margin of 1.1. On the other hand, the classifier B says that the resource should be classified in category #3 because of a margin of 1.2 was returned, but the gap is again slight as compared to the category #2 with a margin of 1.0. The classifier committees would consider all the outputs by adding margins up in order to return a new margin value for the resource upon each category. As a result, committees get the largest margin value for the category #2 with a 2.1, as compared to the 1.8 for the category #3 and 1.7 for the category #1. Hence, both classifiers on their own were wrong classifying this resource, but their prediction criteria were good enough to merge them with other classifiers. The fact that the actual category for the resource was predicted in second place for both classifiers gives rise to the correct classification when using committees.

Next, we show the results of using classifier committees on separate tables for each dataset. Note that the tag-based approach is also included, in order to enable comparing the performance of committees to it.

Table 5.11 on the next page shows the results of using classifier committees on Delicious. The effect of using committees on this dataset is really positive, because all of the combinations considering tags outperform the tag-based classifier. The committees considering only reviews and content may perform worse than tags on their own. Those committees considering the tag-based classifier are the three best. Even though tags positively combine with content and reviews separately,

	Category #1	Category #2	Category #3
Classifier A	1.2	1.1	0.6
Classifier B	0.5	1.0	1.2
Classifier committees	1.7	2.1	1.8

**Table 5.10:** Example of classifier committees, where both classifiers mispredict the category of the resource. One of them predicts category #1, whereas the other predicts category #3. However, it should actually be classified on category #2, which is correctly predicted when adding margins up by using classifier committees.

Delicious - ODP							
Top level							
	600	1400	2200	3000	4000	5000	6000
Tags	.533	.600	.629	.647	.660	.669	.680
Content + Reviews	.554	.604	.627	.642	.651	.660	.670
Content + Tags	.580	.633	.655	.671	.678	.687	.696
Reviews + Tags	.561	.618	.644	.662	.675	.685	.694
Content + Reviews + Tags	.581	.632	.655	.671	.681	.691	.699
Second level							
	600	1400	2200	3000	4000	5000	6000
Tags	.359	.453	.498	.522	.541	.556	.568
Content + Reviews	.382	.450	.486	.505	.522	.538	.547
Content + Tags	.409	.488	.528	.547	.564	.578	.587
Reviews + Tags	.389	.474	.512	.534	.555	.571	.584
Content + Reviews + Tags	.412	.488	.524	.545	.564	.579	.588

**Table 5.11:** Accuracy results of classifier committees for web page classification.

combining all three data sources provides a slight improvement as compared to the other two.

Reviews perform better than content on their own, but the latter performs better when combined with tags. This shows that even though content performs worse, it provides more reliable predictions than reviews, performing better on committees. Nonetheless, relying on all three data sources performs the best in most cases for both levels of the taxonomy.

Table 5.12 on the facing page shows the results of using classifier committees on LibraryThing. These results show the great potential of tags provided by users on this social tagging system. Committees combining data sources not always outperform the sole use of tags. However, combining them with user reviews gives rise to higher performance, especially for the second level classification.

On the other hand, using content on committees yields inferior results. This shows that besides performing worse on its own, content is not good enough in



LibraryThing - DDC							
Top level							
	3000	6000	9000	12000	15000	18000	21000
Tags	.861	.864	.864	.867	.869	.869	.868
Content + Reviews	.778	.803	.814	.821	.823	.827	.830
Content + Tags	.823	.842	.845	.849	.851	.852	.852
Reviews + Tags	.857	.866	.868	.872	.875	.876	.876
Content + Reviews + Tags	.824	.843	.847	.852	.855	.856	.856
Second level							
	3000	6000	9000	12000	15000	18000	21000
Tags	.690	.700	.707	.709	.715	.712	.715
Content + Reviews	.589	.631	.652	.663	.670	.679	.684
Content + Tags	.645	.672	.688	.695	.700	.706	.707
Reviews + Tags	.687	.708	.717	.721	.729	.729	.733
Content + Reviews + Tags	.647	.677	.693	.701	.705	.713	.713
LibraryThing - LCC							
Top level							
	3000	6000	9000	12000	15000	18000	21000
Tags	.853	.857	.856	.861	.861	.857	.861
Content + Reviews	.777	.800	.808	.814	.818	.817	.824
Content + Tags	.787	.806	.815	.819	.824	.821	.830
Reviews + Tags	.831	.845	.853	.856	.861	.859	.864
Content + Reviews + Tags	.791	.811	.820	.826	.831	.827	.838
Second level							
	3000	6000	9000	12000	15000	18000	21000
Tags	.703	.725	.729	.734	.734	.736	.739
Content + Reviews	.600	.648	.674	.690	.705	.704	.719
Content + Tags	.640	.677	.698	.709	.723	.720	.738
Reviews + Tags	.688	.723	.736	.746	.754	.755	.766
Content + Reviews + Tags	.645	.685	.708	.721	.733	.732	.750

**Table 5.12:** Accuracy results of classifier committees for book classification (LibraryThing).

this case to feed classifier committees. Probably, using synopses and editorial reviews as a summary of the content because of the unavailability of the actual content of the book makes it insufficient to get solid results.

Table 5.13 on the next page shows the results of using classifier committees on GoodReads. In this case, tags on their own were not strong enough to reach the results by content or user reviews. However, the committees considering tags perform the best, showing their high reliability when it comes to combining predictions.

As it happened with LibraryThing, content does not seem to be a reliable source for committees. Combining it with reviews and tags yields similar or even worse results than excluding it. Combining both reviews and tags is the best option again for the top level of the taxonomies, as for LibraryThing. Surprisingly, this combination produces results almost as good as using LibraryThing tags, which perform far better on their own. This shows that even though tags from GoodReads are not accurate enough on their own, they provide reliable margins to be considered on committees.

When comparing taxonomies, DDC and LCC, neither GoodReads nor LibraryThing shows any differences as compared to the other, proving that the conclusions are the same regardless of the classification scheme.

Summarizing, tags have shown great potential, not only as a source to classify on their own, but also to provide reliable prediction criteria to take into consideration for combining them with other data sources. Moreover, in some cases like on GoodReads, tags were not good enough on their own, but have shown to be a solid data source when used with classifier committees. Nonetheless, the data source used to combine with tags must be solid enough and provide reliable predictions to get better results. When data sources are selected appropriately, the performance improvement can be considerable. In this regard, we have seen that the synopses and reviews we chose as a summary of the content of books provide inappropriate predictions.

## 5.6 Conclusion

In this chapter, we have carried out a deep experimentation and performed a thorough analysis on the use of social tags as a source to feed resource classifiers. We have compared the performance of using social tags to that by using other data sources like the content or user reviews gathered from social media. The experiments have been applied to the three large-scale social tagging datasets introduced in Chapter 4 on page 59, gathered from tagging sites with different settings and annotated resources, which allow to conclude with more generalistic thoughts. Classification experiments have been realized with annotated web pages over the ODP taxonomy, and annotated books over the DDC and LCC

GoodReads - DDC							
Top level							
	3000	6000	9000	12000	15000	18000	21000
Tags	.745	.747	.754	.757	.757	.757	.756
Content + Reviews	.778	.803	.814	.821	.823	.827	.830
Content + Tags	.797	.822	.831	.837	.838	.844	.845
Reviews + Tags	.820	.847	.857	.865	.867	.872	.874
Content + Reviews + Tags	.806	.831	.842	.849	.851	.854	.857
Second level							
	3000	6000	9000	12000	15000	18000	21000
Tags	.509	.520	.528	.528	.530	.529	.530
Content + Reviews	.589	.631	.652	.663	.670	.679	.684
Content + Tags	.594	.633	.652	.662	.671	.676	.680
Reviews + Tags	.610	.651	.670	.683	.691	.696	.705
Content + Reviews + Tags	.611	.651	.672	.683	.689	.698	.702
GoodReads - LCC							
Top level							
	3000	6000	9000	12000	15000	18000	21000
Tags	.725	.731	.737	.738	.734	.731	.743
Content + Reviews	.777	.800	.808	.814	.818	.817	.824
Content + Tags	.793	.814	.823	.829	.831	.832	.836
Reviews + Tags	.831	.836	.847	.853	.857	.857	.864
Content + Reviews + Tags	.801	.825	.833	.839	.844	.843	.850
Second level							
	3000	6000	9000	12000	15000	18000	21000
Tags	.494	.507	.510	.514	.513	.517	.519
Content + Reviews	.600	.648	.674	.690	.705	.704	.719
Content + Tags	.608	.649	.672	.684	.692	.696	.703
Reviews + Tags	.624	.674	.696	.712	.725	.730	.735
Content + Reviews + Tags	.626	.674	.699	.713	.728	.727	.742

**Table 5.13:** Accuracy results of classifier committees for book classification (GoodReads).

taxonomies. The great potential shown by social tags, for both the top and second levels of taxonomies, can be strengthened by combining the predictions with other data sources. However, not all data sources are strong enough to perform well at combining predictions, so that the selection of data sources should be done appropriately.

Parts of the research in this chapter have been published in [Zubiaga et al. \(2009d\)](#), [Zubiaga et al. \(2009c\)](#) and [Zubiaga et al. \(2011a\)](#).

By means of these experiments, we provided an answer to the following research questions:

#### Research Question 4

*What is the best way of amalgamating users' aggregated annotations on a resource in order to get a single representation for a resource classification task?*

We have shown that it is worthwhile considering all the tags annotated on a resource instead of those in the top that were annotated most. Tags in the top are the most important, and give the main information on the aboutness of resources. However, tags in the tail are helpful to a lesser extent, providing meaningful information and improving the performance of the classifier.

Regarding the weights assigned to those tags when representing a resource, the number of users annotating each tag should be considered in order to get the best results. This is the value that has shown the best results in our experiments. It has outperformed other approaches ignoring weights or considering other data such as the total number of users annotating the resource.

Thereby, the best representation in our experiments is the one that includes all the tags with the values corresponding to the number of users annotating them.

#### Research Question 5

*Despite of the usefulness of social tags for these tasks, is it worthwhile considering their combination with other data sources like the content of the resource as an approach to improve the results even more?*

By means of classifier committees, which combine the predictions by different classifiers, we have shown that tags provide reliable prediction criteria to take into consideration. SVM classifiers not only predict a category, but also assign a weight to each category based on the given resource. These weights, given in the form of margin values, can be used by other classifiers which rely on different data sources. Adding up weights provided by different classifiers can help predict the correct category when a single classifier fails to categorize the resource appropriately. Weights provided by classifiers relying on social tags are especially useful when combining them with results from other classifiers. Nonetheless, not all data sources are helpful for combination in classifier committees, and the selected data source must be solid enough and provide reliable predictions to

outperform the sole use of tags. When data sources are selected appropriately, the performance improvement can be considerable. We have shown that this varies among datasets. For example, with the Delicious dataset, it is important to analyze all three data sources (content, reviews, and tags). However, with the LibraryThing and GoodReads datasets reviews and tags suffice.

**Research Question 6**

*Are social tags also useful and specific enough to classify resources into narrower categories as in deeper levels of hierarchical taxonomies?*

We have analyzed the usefulness of social tags for classification on two different levels of hierarchical taxonomies. Besides broader categories in the top level, we have also explored the classification on narrower categories in the second level. In this regard, social tags have shown to outperform the other data sources on social tagging sites that encourage users to annotate resources (Delicious and LibraryThing). Tags show clear outperformance in these cases, especially on Delicious, where the difference is even more favorable in the second level. This difference is very similar on LibraryThing. Finally, tags from GoodReads do not outperform other data sources at any level because the system does not encourage users to tag books, so that many bookmarks are not annotated.

Our findings provide a different conclusion from that by [Noll and Meinel \(2008a\)](#), where the authors pointed out the hypothesis that social tags were probably useless for deeper levels of taxonomies, and alternative data should be used instead. However, the authors performed just a statistical analysis, and did not confirm the hypothesis with real experiments.



## Analyzing the Distribution of Tags for Resource Classification

*“Statistics will prove anything, even the truth.”*

— Noel Moynihan

In this chapter, we deal with the task of considering the representativity of tags for resource classification within a collection of social annotations on a social tagging system. To the best of our knowledge, no effort has been invested so far on establishing the representativity of tags when it comes to finding the aboutness of resources. In this regard, we explore how the distribution of tags across the three dimensions involved in a social tagging system (namely users, resources and bookmarks) can determine their representativity. To this end, we study and analyze the effectiveness of applying an IDF-like distribution-driven weighting scheme in search of performance improvements in a resource classification task. We define three analogous weighting schemes –IUF, IRF and IBF– which rely on distributions of tags across users, resources and bookmarks, respectively. They have been barely used for social tagging, and their usefulness has not yet been proven.

The chapter is organized as follows. Next, in Section 6.1 we motivate the problem of considering tag distributions as a means to determine the representativity of tags. Then, in Section 6.2 on the facing page we describe the TF-IDF weighting scheme and its use on classical documents collections, and introduce analogous schemes adapted to social tagging systems in Section 6.3 on page 98. We present a set of experiments –tag-based classification, classifier committees, and correlation between weighting measures–, and analyze and study the results in Section 6.4 on page 100. Finally, in Section 6.5 on page 109 we conclude the chapter.

We address the following research questions in this chapter:

**Research Question 7**

*Can we further consider the distribution of tags across the collection so that we can measure the overall representativity of each tag to represent resources?*

**Research Question 8**

*What is the best approach to weigh the representativity of tags in the collection for resource classification?*

## 6.1 Tag Distributions

So far, we have explored the ways of amalgamating great deals of user annotations provided in the form of social tags, in order to find a suitable representation of a resource. We considered the weighting of a tag with respect to the resource where it was annotated, but we did not explore further into the representativity of tags within the whole collection. We have considered that two tags with the same number of users annotating it on a resource have the same representativity for the resource, because they had the same number of annotators and, therefore, they were assigned the same weight. However, they do not strictly have to represent the same representativity.

From a statistical point of view, we believe that the distribution of tags across the whole collection has much to do with the overall representativity of tags. By representativity, we refer to the weight setting how important is a certain tag when it comes to representing a resource for its classification. Accordingly, we believe that a tag that concentrates within a few resources or has been used by a few users is rather representative than a tag present in most resources or used by most users. Even if two tags have the same overall use within the collection, the



way they are distributed across users, resources and bookmarks may determine whether they are focused and precise, or they are spread and imprecise instead.

To this end, a collection-aware weighting scheme like the well-known TF-IDF seems to be a good alternative. We believe it is suitable to determine the representativity of tags considering their distribution across the collection. We found that it had been hardly applied to a social structure like that by tagging systems. Its adaptation from a classical text collection, where the only dimensions are terms and documents, to a collection of bookmarks, where tags spread across users, resources and bookmarks, remains unstudied. Moreover, its usefulness for tag-based resource classification has not yet been explored.

## 6.2 TF-IDF as a Term Weighting Function

TF-IDF is a term weighting function that serves as a statistical measure defining the importance of a word to a document in a collection (Salton et al., 1975; Salton and Buckley, 1988). When computing the TF-IDF value for the term  $i$  within the document  $j$  as a part of a document collection  $D$ , it comprises two underlying measures: (1) the term frequency (TF), i.e., the number of appearances of the term  $i$  within the document  $j$ , and (2) the inverse document frequency (IDF), i.e., the inverse of the number of documents within the whole set of documents  $D$  in which the term  $i$  occurs, which refers to the general importance of the term  $i$  in the collection (see Equation 6.1). The product of these two measures defines the TF-IDF weight of term  $i$  in the document  $j$  (see Equation 6.2).

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|} \quad (6.1)$$

$$tf-idf_{ij} = tf_{ij} \times idf_i \quad (6.2)$$

Integrating the IDF factor allows to rate lower or higher such a term depending on its distribution across the collection. This weighting function yields a higher value when the term  $i$  occurs in a few documents, considering that it is of utmost representativity to those documents. On the other hand, the value will be lower when the term  $i$  occurs in many documents of the collection, considering that it rather spreads across the collection instead of focusing in a few documents. In the latter case, the value becomes null when the term  $i$  occurs in all the documents.

This weighting scheme has been widely used for Information Retrieval, Text Mining and Text Classification, and it is commonly used for term selection tasks. There is controversy on its appropriateness for text classification (Lan et al., 2005; Forman, 2008), since it does not consider the relations between the terms and their appearance in the categories. However, it has shown high effectiveness in

several text classification tasks (Joachims, 1998; Yang and Liu, 1999; Brank et al., 2002; Dumais et al., 1998).

There are some works that study the adaptation of TF-IDF to text and web page classification tasks. They consider the distribution of terms across categories in the training set as a value to determine the representativity of a term. For instance, in Forman (2008) and Lan et al. (2005) the authors compared some feature scoring metrics, including TF-IDF, in a text classification problem using a linear SVM. Each of them proposed a new term weighting function that outperformed TF-IDF in their experiments. Other works, such as Debole and Sebastiani (2003) and Soucy (2005), propose the use of supervised weighting techniques instead of unsupervised ones for text classification tasks.

Even though alternatives to TF-IDF like those mentioned above have been proposed and successfully applied to specific tasks and collections, they have barely been used subsequently. TF-IDF continues to be the most widely used term weighting scheme, and has become a “de facto” standard for document representation.

In this chapter, we rely on TF-IDF as the base weighting to propose analogous schemes adapted to social tagging systems. Even though we could rely on alternatives, our main goals are (1) to perform a study on its adaptability to these structures, and (2) to find out how the settings of social tagging systems affect the resulting tag distributions and thereby the values of such weights. Thus, we will not include any category data in the calculation of the weights.

### 6.3 Tag Weighting Functions Based on Inverse Frequencies

Unlike classical collections of web documents or library catalogs, where the distribution of terms across documents on the collection has been studied, social tagging systems comprise more dimensions to explore into. Besides the distribution of tags across documents or annotated resources, different users set those tags within different bookmarks. These two characteristics are new on social tagging with respect to classical text document collections. Despite this clear difference in the nature of social tagging systems, not enough attention has been paid at analyzing how each of the dimensions –resources, users and bookmarks– affects tag distributions and, therefore, establishing tag relevances.

TF-IDF has widely been applied to text collections, and has proven to be beneficial for a large number of tasks. Text collections are mainly made up by terms written by the authors, though, and the appropriateness of using a similar approach for a collection made up by tags annotated by users other than the authors on a social environment is not clear.

Next, we introduce three tag weighting approaches, taking the classical TF-IDF approach to the social tagging scenario, and adapting it to rely on resources, users and bookmarks. These three dimensions suggest the definition of that many tag weighting functions considering inverse resource frequency (IRF), inverse user frequency (IUF), and inverse bookmark frequency (IBF) values, respectively. These three approaches follow the same function for the tag  $i$  within the resource  $j$  (see Equation 6.3).

$$TF\text{-}IxF_{ij} = tf_{ij} \cdot ix_f \quad (6.3)$$

where  $tf_{ij}$  is the number of occurrences of the tag  $i$  in the resource  $j$ , and  $ix_f$  is the inverse frequency function considered in each case,  $irf$ ,  $iuf$  or  $ibf$ , thus  $x$  being  $r$ ,  $u$ , or  $b$ .

### 6.3.1 TF-IRF

This is the application of the TF-IDF approach to a social tagging system with annotated resources, considering that resources are analogous to documents in this case. Tags that are widely spread across resources are penalized with low weights and, vice versa, tags within fewer resources are considered relevant with a higher weight. Thus, the function outputs the logarithm of the total number of resources divided by the number of resources in which the tag is present (see Equation 6.4).

$$irf_i = \log \frac{|R|}{|\{r : t_i \in R\}|} \quad (6.4)$$

It has previously been used in a few works in the social tagging literature, even though they usually referred to this approach as TF-IDF. [Angelova et al. \(2008\)](#) rely on this measure to infer similarity of tags by creating a tag graph, weighting the TF-IDF value of each user to a tag. [Shepitsen et al. \(2008\)](#) and [Liang et al. \(2010\)](#) use this measure to represent the resources in a recommendation system where resources are recommended to users. The latter concluded that although both TF-IDF and TF have identical trends, the former provides superior results in their recommendation task. Likewise, [Ramage et al. \(2009\)](#) compared TF-IDF and TF for clustering web pages, and showed a superiority for the former. However, they did not pay attention at the effect of tag distributions on these weightings, and they showed the usefulness of TF-IDF just for a specific case. [Li et al. \(2008\)](#) create tag vectors using TF-IDF to compute the similarity between two documents annotated on Delicious. They assumed this weighting measure, and they did not pay attention at whether or not it was appropriate.

### 6.3.2 TF-IUF

As a new dimension present in social tagging systems, the number of users using each of the tags could also be significant to know whether a tag is representative to a collection of resources. Thus, we consider that a tag used by many users is not as representative as a tag that fewer users are utilizing (see Equation 6.5).

$$iuf_i = \log \frac{|U|}{|\{u : t_i \in U\}|} \quad (6.5)$$

This function was inferred from a previous application to a collaborative filtering system Breese et al. (1998). With the aim of recommending resources to users, Diederich and Iofciu (2006) and Liang et al. (2010) rely on the IUF for discovering similarities among users. The latter use both IUF and IRF to represent users and resources, respectively, but no comparison is performed among their characteristics. In Abbasi et al. (2009), TF-IUF is used along with TF-IRF over Flickr tags and user groups for finding landmark photos. They concluded that their approach was effective to find landmark photos on Flickr, but they did not study whether or not relying on those weighting measures was appropriate.

### 6.3.3 TF-IBF

This is a similar inverse weighting function relying on the third dimension in which tags are distributed: bookmarks. This function considers that a tag that has been used in many bookmarks is not as relevant to represent a resource as others that have been assigned to fewer bookmarks (see Equation 6.6).

$$ibf_i = \log \frac{|B|}{|\{b : t_i \in B\}|} \quad (6.6)$$

To the best of our knowledge, this tag weighting scheme has never been used so far. Even though all three frequencies can somehow be related, there are substantial differences among them. A tag used by many users can spread across many resources, or it can just congregate in a few resources. Likewise, this factor might affect the number of bookmarks.

## 6.4 Experiments

Next, we present the classification experiments that enable (1) to analyze how each of the proposed tag weighting functions contributes to the classification of annotated resources, as well as (2) to discover whether either of the inverse tag weighting approaches outperforms the baseline relying only on the tag frequency (TF). In order to further analyze their usefulness and suitability, we also experimented on their performance when applied to classifier committees. Finally, we analyze the correlation between the different tag weighting functions.

### 6.4.1 Tag-based Classification

The first experiment focuses on evaluating the usefulness of tag weighting functions for a resource classification task. We perform this evaluation by comparing tag-based representations by using each of the three weighting functions –TF-IRF, TF-IUF and TF-IBF– and the absence of distributional weighting functions (TF). Note that the latter is the same as the FTA-based weighted approach we concluded as the best representation in Chapter 5 on page 75, and it is thus the up-to-now outperforming approach. This experiment uses an SVM with the same settings as those defined in previous Chapter (see Section 5.3 on page 79). We show the results for all three datasets, and 4 different representations, including the three weighting measures and TF.

Delicious - ODP							
Top level							
	600	1400	2200	3000	4000	5000	6000
TF	.533	.600	.629	.647	.660	.669	.680
TF-IRF	.516	.571	.593	.607	.619	.631	.639
TF-IBF	.519	.573	.596	.611	.622	.633	.641
TF-IUF	.528	.580	.607	.625	.636	.653	.661
Second level							
	600	1400	2200	3000	4000	5000	6000
TF	.359	.453	.498	.522	.541	.556	.568
TF-IRF	.344	.424	.463	.486	.506	.518	.529
TF-IBF	.348	.429	.467	.489	.509	.520	.532
TF-IUF	.358	.437	.478	.502	.523	.541	.555

**Table 6.1:** Accuracy results of tag-based web page classification using weighting schemes.

Table 6.1 shows the results of using tag weighting functions on Delicious. It can be seen that the use of inverse weighting functions is not useful in this case. In the contrary, their use harms the performance of the classifier, yielding inferior results than those obtained by the TF approach not considering weighting functions. Going further into the analysis of the performance of representations relying on weighting functions, the results show that IUF gets the best results among them, followed by IBF, and then IRF. This happens for both top and second levels of the taxonomy in a similar manner.

Our conjecture about this is that resource-based tag suggestions provided by Delicious are not helpful to this end. We have already shown in Chapter 4 on page 59 that such a feature alters the structure of the folksonomy on Delicious. It makes the top tags become even more popular and it alters the natural distribution of tags. Thus, such a forced distribution of tags produces weights that score lower performances. Moreover, the fact that IUF is the best weighting

function in this case shows the importance of users who make their own choices instead of relying on suggestions. That is, users who are able to choose their own tags and differ from those relying on suggestion-based annotations give rise to higher weights for their seldom tags. When users rely on suggestions, it does not make any difference on the IRF values of tags, because the frequency remains unchanged. This difference is also little for IBF values. However, it makes a big difference on IUF values, because those suggestions increase the user frequencies of tags and thus reduce IUF values. Accordingly, users who make their own choices yield higher IUF values because it is likely that their tags are not being used that many times. Probably, IUF would perform better than TF if there were fewer users who rely on system suggestions, and hence more users providing their own tags instead.

LibraryThing - DDC							
Top level							
	3000	6000	9000	12000	15000	18000	21000
TF	.861	.864	.864	.867	.869	.869	.868
TF-IRF	.877	.889	.894	.897	.900	.902	.902
TF-IBF	.877	.889	.894	.897	.900	.903	.904
TF-IUF	.881	.891	.895	.897	.899	.901	.900
Second level							
	3000	6000	9000	12000	15000	18000	21000
TF	.690	.700	.707	.709	.715	.712	.715
TF-IRF	.723	.750	.762	.768	.774	.777	.780
TF-IBF	.723	.751	.763	.770	.775	.779	.781
TF-IUF	.729	.751	.761	.766	.771	.771	.776
LibraryThing - LCC							
Top level							
	3000	6000	9000	12000	15000	18000	21000
TF	.853	.857	.856	.861	.861	.857	.861
TF-IRF	.867	.883	.887	.893	.895	.894	.897
TF-IBF	.867	.883	.888	.893	.896	.895	.898
TF-IUF	.871	.882	.885	.892	.893	.892	.894
Second level							
	3000	6000	9000	12000	15000	18000	21000
TF	.703	.725	.729	.734	.734	.736	.739
TF-IRF	.751	.780	.793	.803	.804	.809	.814
TF-IBF	.751	.781	.796	.805	.806	.811	.818
TF-IUF	.754	.780	.790	.798	.800	.803	.807

**Table 6.2:** Accuracy results of tag-based book classification using weighting schemes (LibraryThing).

Table 6.2 shows the results of using tag weighting functions on LibraryThing over DDC and LCC schemes. In this case, all the inverse weighting functions

are clearly superior to TF, since the former always outperform the latter. Even though the outperformance is much larger for the second level, the superiority of weighting functions is clear for both levels. This shows that the studied inverse weighting functions can be really useful for folksonomies created in the absence of suggestions. Inverse tag weighting functions have successfully set suitable weights towards a definition of the representativity of tags in this case, in contrast to Delicious.

Among the tag weighting functions, all of them perform similarly, and no clear outperformances can be seen in these results. However, IBF seems to provide slightly better results than the other two approaches, followed by IRF. IUF is the worst function in this case, suggesting that the number of users choosing each tag is not the most relevant feature when there are no suggestions.

GoodReads - DDC							
Top level							
	3000	6000	9000	12000	15000	18000	21000
TF	.745	.747	.754	.757	.757	.757	.756
TF-IRF	.800	.808	.813	.817	.816	.817	.816
TF-IBF	.800	.809	.814	.817	.817	.818	.818
TF-IUF	.797	.805	.810	.814	.813	.814	.814
Second level							
	3000	6000	9000	12000	15000	18000	21000
TF	.509	.520	.528	.528	.530	.529	.530
TF-IRF	.579	.599	.609	.612	.617	.619	.621
TF-IBF	.583	.602	.614	.618	.624	.626	.628
TF-IUF	.578	.598	.609	.613	.619	.620	.623
GoodReads - LCC							
Top level							
	3000	6000	9000	12000	15000	18000	21000
TF	.725	.731	.737	.738	.734	.731	.743
TF-IRF	.781	.792	.797	.801	.802	.799	.804
TF-IBF	.781	.792	.797	.803	.802	.800	.805
TF-IUF	.776	.788	.792	.797	.797	.794	.800
Second level							
	3000	6000	9000	12000	15000	18000	21000
TF	.494	.507	.510	.514	.513	.517	.519
TF-IRF	.578	.599	.608	.617	.618	.622	.627
TF-IBF	.582	.605	.615	.625	.625	.628	.634
TF-IUF	.576	.600	.610	.619	.620	.623	.628

**Table 6.3:** Accuracy results of tag-based book classification using weighting schemes (GoodReads).

Table 6.3 shows the results of using inverse tag weighting functions over DDC and LCC schemes on GoodReads. Similar to LibraryThing, tag weighting func-

tions clearly outperform the sole use of TF. Moreover, these outperformances are even superior than for LibraryThing. As on LibraryThing, IBF performs the best among the weighting functions, followed by IRF, and then IUF.

Even though there are also system suggestions on GoodReads, they rely on tags previously used by the user, i.e., their personomy, and thus these suggestions can only be applied to different bookmarks and resources. Thereby, those users who tend to choose new tags instead of reusing tags from their personomy are yielding more natural bookmark frequencies. This affects and helps IBF perform better, but has no impact on IUF, as it is not altered by personomy-based suggestions. This shows that the effect of personomy-based suggestions is much smaller, and it affects to a lower extent or does not almost affect the distribution of tags, because suggestions do not spread to the users. Accordingly, the studied tag weighting functions perform well when this type of suggestion exists. On both LibraryThing and GoodReads, the results for the different classification schemes, DDC and LCC, are comparable and show a similar trend.

Summarizing, results show that the studied inverse tag weighting functions can be really useful for determining the representativity of each tag within the collection. However, folksonomies can suffer from resource-based tag suggestions, transforming the structure and distributions of folksonomies. This transformation can even be harmful for the definition of tag weighting functions, and can bring about worse performance results than simply relying on TF, as happened on Delicious. Otherwise, in the absence of resource-based tag suggestions, the use of tag weighting functions contribute in a positive manner to the performance of the classifier.

Comparing the results scored by tag weighting functions, it can be seen that IBF is always slightly better than IRF. The former is more detailed than the latter, because it considers the exact number of appearances of the tag besides the number of resources it appears in. Actually, IBF is the best approach for both LibraryThing and GoodReads, where there are no suggestions, or suggestions rely on user's personomy. When these suggestions rely on tags previously annotated by others to the resource, as on Delicious, IUF performs better than the other two weighting functions, showing the relevance of the ability of users to dismiss suggestions. However, even IUF is not able to outperform TF in this case.

## 6.4.2 Revisiting Classifier Committees

Apart from the results scored using tag weighting functions and the comparison of their performance to that by relying on TF, it is interesting to analyze their appropriateness to combine with other data sources. As we did in [Chapter 5 on page 75](#), we use classifier committees to evaluate the ability of the approaches using tag weighting functions to be combined with content and/or reviews, and



improve even more the performance of the classifier. This time, we rely on the best committees for each datasets, i.e., the triple combination of tags, content and reviews for Delicious, and the double combination of tags and reviews for LibraryThing and GoodReads. We run them using the 4 different weightings for tags: TF, TF-IBF, TF-IRF, and TF-IUF, i.e., those compared in the previous section as well. By using classifier committees upon these weightings, we aim at analyzing how well they perform not only on their own, but also providing their prediction criteria when combining with other data sources.

Delicious - ODP							
Top level							
	600	1400	2200	3000	4000	5000	6000
TF	.581	.632	.655	.671	.681	.691	.699
TF-IRF	.576	.629	.653	.669	.680	.690	.697
TF-IBF	.576	.630	.653	.670	.680	.690	.698
TF-IUF	.576	.631	.654	.672	.682	.692	.700
Second level							
	600	1400	2200	3000	4000	5000	6000
TF	.412	.488	.524	.545	.564	.579	.588
TF-IRF	.406	.485	.523	.546	.566	.580	.592
TF-IBF	.407	.486	.525	.548	.566	.580	.592
TF-IUF	.408	.488	.526	.548	.569	.584	.595

**Table 6.4:** Accuracy results of classifier committees for web page classification using weighting schemes.

Table 6.4 shows the classification results of the approaches considering inverse tag weighting functions on classifier committees for Delicious. Even though inverse tag weighting functions were not useful to improve the performance of the tag-based classifier reducing its overall accuracy, they seem to provide better decisions to be combined with other data sources. The predictions and margins outputted by all three approaches using inverse weighting functions yield slightly better results on the classification, especially when it comes to second level classification. Thereby, inverse tag weighting functions are useful for Delicious when their outputs are applied on classifier committees along with content and reviews. The unsuitability of tag weighting functions on their own gets fixed by the use of classifier committees. However, the little outperformance by tag weighting functions when using committees is almost irrelevant as compared to the TF-based committees. This outperformance is slightly clearer for largest training sets upon the second level classification.

Table 6.5 on the next page shows the classification results of the approaches considering tag weighting functions on classifier committees for LibraryThing over DDC and LCC schemes. In this case, inverse tag weighting functions are also

LibraryThing - DDC							
Top level							
	3000	6000	9000	12000	15000	18000	21000
TF	.857	.866	.868	.872	.875	.876	.876
TF-IRF	.864	.882	.886	.890	<b>.894</b>	<b>.897</b>	.897
TF-IBF	<b>.865</b>	<b>.883</b>	<b>.887</b>	<b>.891</b>	<b>.894</b>	<b>.897</b>	<b>.898</b>
TF-IUF	<b>.865</b>	<b>.883</b>	.886	.889	.892	.894	.895
Second level							
	3000	6000	9000	12000	15000	18000	21000
TF	.687	.708	.717	.721	.729	.729	.733
TF-IRF	.709	<b>.742</b>	.754	.765	.770	.773	.778
TF-IBF	.710	<b>.742</b>	<b>.756</b>	<b>.767</b>	<b>.772</b>	<b>.776</b>	<b>.780</b>
TF-IUF	<b>.712</b>	.741	.752	.763	.767	.769	.775
LibraryThing - LCC							
Top level							
	3000	6000	9000	12000	15000	18000	21000
TF	.831	.845	.853	.856	.861	.859	.864
TF-IRF	.849	.869	.876	.880	.887	.885	.890
TF-IBF	.851	<b>.871</b>	<b>.879</b>	<b>.882</b>	<b>.888</b>	<b>.887</b>	<b>.892</b>
TF-IUF	<b>.852</b>	.869	.875	.880	.886	.885	.888
Second level							
	3000	6000	9000	12000	15000	18000	21000
TF	.688	.723	.736	.746	.754	.755	.766
TF-IRF	.712	.750	.770	.782	.789	.793	.803
TF-IBF	.717	<b>.755</b>	<b>.773</b>	<b>.786</b>	<b>.792</b>	<b>.797</b>	<b>.806</b>
TF-IUF	<b>.719</b>	.754	.770	.781	.788	.792	.801

**Table 6.5:** Accuracy results of classifier committees for book classification using weighting schemes (LibraryThing).

useful when applied to classifier committees when compared to the TF-based one. Those classifier committees including tag weighting functions produce clearly better results than the committee using TF. This performance improvement is positive for both levels, but it is larger for the second level. However, those approaches using tag-based representations with tag weighting functions perform better on their own, without considering committees (see Table 6.2 on page 102). That is, it is better to use the classifier based on tags on their own, without including the predictions by the classifier using reviews. This means that predictions with tag weighting functions are good enough to work on their own, and it is better to ignore the other data source, i.e., reviews, which cannot catch up with the performance of tags and harm the overall performance.

Table 6.6 on the next page shows the classification results of the approaches considering tag weighting functions on classifier committees for GoodReads over DDC and LCC schemes. The main conclusions drawn from these results are very

GoodReads - DDC							
Top level							
	3000	6000	9000	12000	15000	18000	21000
TF	.820	.847	.857	.865	.867	.872	.874
TF-IRF	.835	.859	.867	.874	.877	.881	.884
TF-IBF	.837	.861	.868	.876	.878	.882	.885
TF-IUF	.834	.858	.866	.873	.876	.881	.883
Second level							
	3000	6000	9000	12000	15000	18000	21000
TF	.610	.651	.670	.683	.691	.696	.705
TF-IRF	.637	.676	.693	.707	.716	.719	.726
TF-IBF	.642	.681	.697	.711	.719	.723	.730
TF-IUF	.638	.677	.694	.708	.717	.722	.727
GoodReads - LCC							
Top level							
	3000	6000	9000	12000	15000	18000	21000
TF	.831	.836	.847	.853	.857	.857	.864
TF-IRF	.826	.846	.856	.861	.866	.864	.870
TF-IBF	.829	.848	.858	.863	.868	.866	.871
TF-IUF	.826	.845	.856	.860	.866	.864	.869
Second level							
	3000	6000	9000	12000	15000	18000	21000
TF	.624	.674	.696	.712	.725	.730	.735
TF-IRF	.647	.697	.716	.732	.742	.748	.757
TF-IBF	.651	.700	.720	.736	.746	.751	.759
TF-IUF	.648	.698	.718	.733	.744	.749	.757

**Table 6.6:** Accuracy results of classifier committees for book classification using weighting schemes (GoodReads).

similar to those on LibraryThing. Again, committees relying on tag weighting functions perform clearly better than that relying on TF, especially for the second level. However, it is better to ignore the other data source, i.e., reviews, since the results by tags on their own are good enough and cannot be improved by combining them (see Table 6.3 on page 103).

Again, using classifier committees obtains comparable results with very similar trends for both book taxonomies, DDC and LCC.

Summarizing the results for all three datasets, the use of tag weighting functions has shown to be helpful in all cases as compared to TF when it comes to combining them with other data sources using classifier committees. However, it is better to rely only on the tag-based classifier for both LibraryThing and GoodReads, which score good results on their own, and they get harmed when combined with other data sources. In the case of Delicious, on the other hand, the use of classifier committees for approaches relying on tag weighting functions

perform better results than using tags on their own, and than the committees relying on TF. Nonetheless, the latter performs just slightly worse, and their results are very similar, suggesting that any of them could be used to perform the task.

### 6.4.3 Correlation between Tag Weighting Functions

All three inverse tag weighting functions consider the distribution of tags across different dimensions. The values given by these three functions could correlate or not depending on the behavior of users, e.g., if many tags annotated by a large number of users congregate into the same resource, correlation between IUF and IRF would be lower than if each of the users annotate those tags in different resources. Thus, analyzing whether these three values correlate is of utmost importance.

Table 6.7 shows correlation values among tag weighting schemes. The correlation values between each pair of functions are shown in each row, for both the Pearson and Spearman correlation coefficients. Note that the latter considers the rank inferred from tag weights, whereas the former considers the values to compute correlations. Both correlation values range from -1 to 1. The closer is this value to 0, the less correlation exists among the compared sets and, thus, the more independent they are.

	Delicious		LibraryThing		GoodReads	
	$r$	$\rho$	$r$	$\rho$	$r$	$\rho$
IRF-IUF	.763	.657	.679	.603	.529	.421
IRF-IBF	.991	.990	.989	.981	.997	.998
IUF-IBF	.780	.677	.720	.630	.556	.436

**Table 6.7:** Pearson ( $r$ ) and Spearman ( $\rho$ ) correlation coefficients.

Correlation values show that there is a high dependence among IBF and IRF values. Both seem to be fully dependent and, thus, that is why these two approaches achieve very similar results. The correlation decreases when IUF is considered, so that it seems to be more independent to the rest. This independence is clearest for GoodReads, and intermediate for LibraryThing, but Delicious shows the highest dependence of IUF with respect to the other two values. The main reason for the clear independence of IUF on GoodReads is that users are suggested by the system with tags in their personomy, so that they easily spread tags on bookmarks and resources, keeping the user frequency unchanged. As LibraryThing and Delicious do not have this feature, IUF correlates with the others to a greater extent.

## 6.5 Conclusion

In this chapter, we have studied and analyzed the application of tag weighting functions based on the classical IDF scheme for the resource classification task on the three large-scale datasets introduced in Chapter 4 on page 59. We have considered the distributions of tags across users, resources and bookmarks to generate three variations of such weighting, namely IBF, IRF and IUF. We have performed classification experiments by considering their results on their own, and by combining them with other data sources using classifier committees. We have analyzed their results by taking into account the settings of each social tagging system, and how it affects the distribution of tags of their underlying folksonomies.

With these experiments, we have given an answer to the following research questions:

### Research Question 7

*Can we further consider the distribution of tags across the collection so that we can measure the overall representativity of each tag to represent resources?*

We have analyzed the suitability of IDF-like weighting functions to define the representativity of tags, which consider the distribution of tags through the whole collection of resources. Our experiments have shown that these functions helps improve performance of a resource classification task. However, we have shown that the settings of the social tagging system have an effect on those distributions. Resource-based tag suggestions have shown to influence the structure of folksonomies greatly. Suggesting tags based on previous annotations of others on the resource causes a very different tag distribution, which in turn, affects the results of the weighting function. When a system enables the resource-based tag suggestions, the use of tag weighting functions performs worse, and combining with other data sources is required to improve performance; this method can even outperform the TF-based approach.

For our classification experiments, we have found that IDF-like weighting functions clearly outperform the TF approach when resource-based tag suggestions are not enabled, i.e., on LibraryThing and GoodReads, both when used on their own, or when combined with other data sources. We found it better to consider just the tag-based approach, without combining them with other data sources, since it provides superior results, which cannot be improved by combining them with other predictions.

### Research Question 8

*What is the best approach to weigh the representativity of tags in the collection for resource classification?*

Among the studied weighting functions, the one relying on bookmark frequencies has shown to be the best when there are no resource-based tag suggestions. In these cases, IBF performs the best, followed by IRF, and IUF. All of them clearly outperform TF, when both used on their own, and combined with other data sources using classifier committees.

On the other hand, when the social tagging system suggests tags to the user relying on the resource itself, IUF performs better than the others. IUF performs better than IBF and IRF, because of the importance of the ability of users to choose their own tags without relying on suggestions from these systems. Even though IUF does not outperform TF when used on its own, combining it with other data sources produces the best approach. However, it is only slightly better than the committees relying on TF, and any of them can be used to score similar results.

## Analyzing the Behavior of Users for Classification

*“Always imitate the behavior of the winners when you lose.”*

— George Meredith

In this chapter, we explore the behavior of users on social tagging systems. Earlier works have suggested and shown that users of these systems follow different goals, and they tag resources for a certain purpose. Several classifications have been proposed to discriminate user behavior. Specifically, we consider one of those classifications of behavioral purposes. Such classification splits user behavior into two goals: (a) users who aim at maintaining an organizational structure of the resources for later browsing, so-called Categorizers, and (b) users who rather provide detailed descriptions for later search, so-called Describers. These two user behaviors yield different personomy structures, i.e., they follow a different tagging pattern, which produces different tag selections from each other.

Such a classification of users has been previously experimented, and has shown its effectiveness to discover users who rather describe resources, i.e., Describers. However, the appropriateness of Categorizers for a resource classification task has not yet been studied. Upon this, we set out the study of the suitability of users who fit such behavior, by performing a set of resource classification and descriptiveness experiments. To this end, we split the whole set of users into smaller subsets of utmost Categorizers and Describers. We explore how each subset of users better fits the classification or descriptiveness task.

This chapter is organized as follows. Next, in Section 7.1 we briefly summarize the research work found so far on the user motivation to tagging, and motivate its interest towards our work on resource classification. In Section 7.2 on the facing page we detail in more depth one of those classifications of user behavior, which separates Categorizers from Describers. Then, in Section 7.3 on page 115 we present the settings of our resource classification experiments enhanced by the detection of user behavior, and present their results in Section 7.4 on page 119. Finally, we conclude the chapter in Section 7.5 on page 123.

We address the following research questions in this chapter:

**Research Question 9**

*Can we discriminate different user profiles so that we can find a subset of users who provide annotations that better fit a classification scheme?*

**Research Question 10**

*What are the features that identify a user as a good contributor to the resource classification?*

## 7.1 User Behavior on Social Tagging Systems

It has been suggested that not all the users contributing on social tagging systems are motivated by the same goal for annotating resources. Depending on their annotations, several works propose different classifications of user behavior (Körner et al., 2010b). Some of them focus on detecting the types of tags provided by users. For instance, early works such as Golder and Huberman (2006) and Sen et al. (2006) propose the existence of several tag types. On the other hand, others have suggested discriminating user behavior by their annotations. In this regard, works such as Marlow et al. (2006b), Heckner et al. (2009), Nov et al. (2009) and Strohmaier et al. (2010a) propose differentiating users by their motivation for tagging resources.

As a classification of user behavior that matches our requirements, we focus on the latter by Strohmaier et al. (2010a). In this work, the authors propose differentiating two kinds of user behavior: Categorizers, who rather organize resources, and Describers, who rather define the contents of resources. It seems reasonable that users so-called Categorizers may provide annotations that better fit the resource classification task than Describers. Next, we detail in more depth these two types of users.



	Categorizer	Describer
Goal of Tagging	later browsing	later retrieval
Change of Tag Vocabulary	costly	cheap
Size of Tag Vocabulary	limited	open
Tags	subjective	objective

Table 7.1: Characteristics of Categorizers and Describers.

## 7.2 Categorizers vs Describers

The approach we consider for discriminating users by their behavior has been introduced and experimented in earlier works (Körner et al., 2010a,b; Körner, 2009). They consider the existence of two major tagging motivations on social tagging systems: Categorizers and Describers.

Early works such as Marlow et al. (2006b); Hammond et al. (2005a) and Heckner et al. (2009) suggest that a distinction between at least two types of user motivations for tagging is interesting: on one hand, users can be motivated by categorization (in the following, *Categorizers*). These users view tagging as a means to categorize resources according to some (shared or personal) high-level conceptualizations. They typically use a rather elaborated tag set to construct and maintain a navigational aid to the resources for later browsing. On the other hand, users who are motivated by description (so-called *Describers*) view tagging as a means to accurately and precisely detail resources. These users tag because they want to produce annotations that are useful for later search and retrieval. The development of a personal, consistent ontology to navigate across their resources is not their intuition. Table 7.1 gives an overview of characteristics of the two different types of users, based on Körner (2009).

### 7.2.1 Measures

We use three different measures to differentiate users into Categorizers and Describers: Tags Per Post (TPP), Tag Resource Ratio (TRR), and Orphan Ratio (ORPHAN). Additional measures are shown in Körner et al. (2010b), but due to the high correlation with the others, we limited our efforts to the ones above. These measures rely on two features of user behavior: verbosity, which measures the number of tags a user tends to use when annotating, and diversity, which measures the extent to which users are using new tags that were not previously applied by themselves. It is worthwhile noting that these measures provide one value for each user. The measure corresponding to each user is thus computed by considering the characteristics of their bookmarks and the attached tag assignments. The resulting measures are then ranked in a list along with the rest of the users. This list makes possible inferring the extent to which a user is rather a

Categorizer or a Descriptor.

#### 7.2.1.1 Tags per Post (TPP)

As a Descriptor would focus on describing their resources in a very detailed manner, the number of tags used to annotate each resource can be taken into account as an indicator to identify the motivation of the analyzed user. The *tags per post* measure (short *TPP*) captures this by dividing the number of all tag assignments of a user by the number of resources (see Equation 7.1).  $T_{ur}$  is the number of tags annotated by a user  $u$  on a resource  $r$ , and  $R_u$  is the number of resources of the user. The more tags a user utilizes to annotate the resources, the more likely they are a Descriptor, reflecting it in a higher TPP score.

$$TPP(u) = \frac{\sum_r |T_{ur}|}{|R_u|} \quad (7.1)$$

This measure relies on the verbosity of users, as it computes the average number of tags they assigned to bookmarks.

#### 7.2.1.2 Orphan Ratio (ORPHAN)

Since Describers do not have a fixed vocabulary and freely choose tags to describe their resources in a detailed manner, they would not focus on reusing tags. This factor is analyzed in the *orphan ratio* (short *ORPHAN*). This measure relates the number of seldom used tags to the total number of tags. Equation 7.2 shows how seldom used tags are defined by the individual tagging style of a user. In this equation,  $t_{max}$  denotes the most frequent tag of the user. Equation 7.3 shows the calculation of the final measure where  $T_u^o$  are seldom used tags and  $T_u$  are all tags of the given user. Users with more seldom tags yield a higher orphan ratio, and they are more likely to be Describers.

$$n = \left\lceil \frac{|R(t_{max})|}{100} \right\rceil \quad (7.2)$$

$$ORPHAN(u) = \frac{|T_u^o|}{|T_u|}, T_u^o = \{t \mid |R(t)| \leq n\} \quad (7.3)$$

By measuring whether users frequently use the same tags or rather rely on new ones, the ORPHAN ratio considers their diversity.

#### 7.2.1.3 Tag Resource Ratio (TRR)

The *tag resource ratio* (short *TRR*) relates the number of tags of a user (i.e., the size of their vocabulary) to the total number of annotated resources (see Equation

7.4). A typical Categorizer would use a small number of tags as compared to the number of resources and would therefore score a low TRR value.

$$TRR(u) = \frac{|T_u|}{|R_u|} \quad (7.4)$$

This measure relies on both verbosity, because users who use more tags in each bookmark would usually result in a higher TRR value, and diversity, as those who frequently use new tags will have a larger vocabulary. Nonetheless, the latter has a higher impact in this case, since the former could be altered by verbose users who tend to reuse tags.

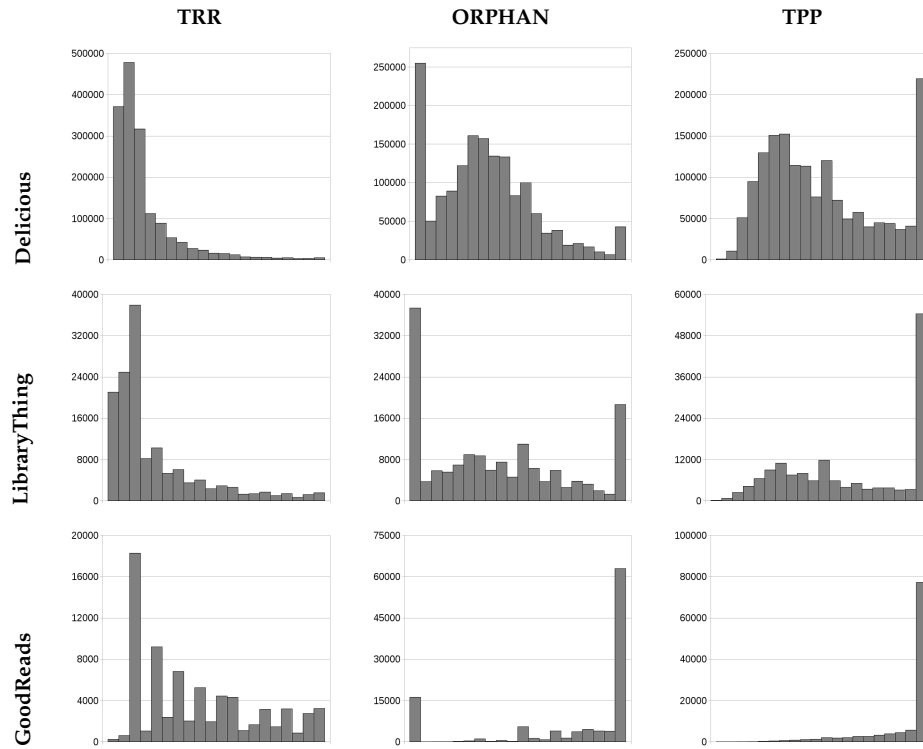
### 7.3 Calculation of Measures and Experiment Settings

Users of each social tagging site have their own weights for each of the three measures above. Thus, we computed TPP, ORPHAN and TRR values for each user. This way, we are able to generate three ranked lists of users for each site. In these rankings, Categorizers rank high, whereas Describers rank low (this is arbitrary and could be inverted as well). From these lists, we can select a subset of users in the top as Categorizers, and another subset in the tail as Describers. Both sets should have the same size in order to compare them.

Our main goal is to conclude whether these measures can discriminate Categorizers in such a way that they perform better than Describers on a resource classification task. However, we also perform experiments measuring the descriptiveness of users' tags in order to conclude whether Describers perform better to that end. With the subsets of Categorizers and Describers defined above, we perform classification and descriptiveness experiments to know how suitable they are for each of the tasks.

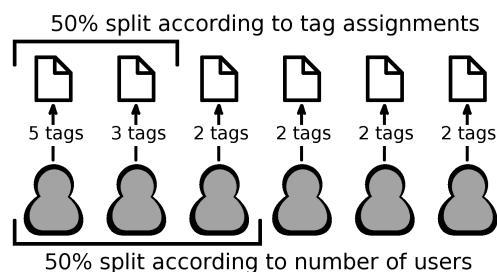
Table 7.2 on the following page shows the distribution of the three measures we calculated for users on the three datasets. The X axis represents quantiles of values, whereas Y axis represents the number of users belonging to each quantile. Note that the values themselves are not relevant, but just allows us to rank each user and analyze where they fall in the distribution of all weights. On one hand, the TRR measure follows a similar distribution for all three datasets. On the other hand, for the other two measures, the distributions show there are lots of extreme users for LibraryThing and GoodReads: (1) the ORPHAN measure shows both many extreme Categorizers and extreme Describers, and (2) the TPP measure discriminates a large set of extreme Categorizers, but almost no extreme Describers. These distributions change drastically on Delicious, though. For this dataset, there are many users who have middle values, and who are not that clearly discriminated as Describers or Categorizers.

To choose the sets of users with which we perform the experiments, we split



**Table 7.2:** Distribution histograms of the three measures (TRR, ORPHAN and TPP) for the three datasets. X axis represents the quantiles of values, whereas Y axis represents the number of users in each quantile.

the ranked lists by getting some of the top and bottom users. Choosing fixed percents of users would be unfair, though. Some users are likely to be more verbose, by definition of some measures, and they usually provide much more tag assignments than others. Thus, we split the users according to the percent of tag assignments they provide<sup>1</sup>. This enables a fairer split of the users, with the same amount of data, e.g., a 10% split ensures that both sets include 10% of all tag assignments, but the number of users differs among them. Figure 7.1 shows an example of how splitting by number of tag assignments can differ from splitting by number of users. We split the user sets into smaller subsets of users ranging from 10% to 100%, with a step size of 10%.



**Figure 7.1:** Example of a 50% segment, selected based on tag assignments or number of users. Splitting by number of users would be unfair, since it may yield bigger amounts of data.

### 7.3.1 Tag-based classification

For the tag-based classification, we represent the resources by aggregating annotations provided by the users within the considered subset of Categorizers or Describers. This creates reduced tagging data for each resource. With these reduced representations, we feed the multiclass SVM classifier defined in Chapter 3 on page 47, and calculate their performance by measuring the accuracy of their predictions. This enables comparing same percents of tag assignments by Categorizers and Describers, in order to analyze whether the former outperform the latter.

### 7.3.2 Descriptiveness of Tags

To compute the extent to which a subset of users is providing descriptive tags, we compare their tags to the descriptive data of resource. These descriptive data include:

<sup>1</sup>We define each of the tags annotated in a bookmark as a tag assignment. Thus, a bookmark has as many tag assignments as tags has the user annotated on it.

- The textual content of the web pages, as well as user reviews for the Delicious dataset.
- Synopses, user reviews and editorial reviews for the book datasets, i.e., LibraryThing and GoodReads.

In the first step, we merge all these data into a single text for each resource. Accordingly, we get a single text comprising all descriptive data for each resource. After this, we compute the frequencies of each term (TF) in the texts, so that we can create a vector for each resource, where each of the dimensions in the vectors belongs to a term. On the other hand, for each selection of users, we create the vectors of tags for each resource, with the annotations of those users. This way, we have the reference descriptive vectors as well as the tag vectors we want to compare to them.

There are several measures that could compute the similarity between a tag vector ( $T$ ) and a reference vector ( $R$ ) for a given resource  $r$ . They tend to be correlated, though. Regardless of the values given by the measures, we are interested in getting comparable values towards a way to determine whether a tag set resembles to a greater or lesser extent than another set. Thus, as a well-known and robust measure for this, we compute the cosine similarity between the vectors (see Equation 7.5).

$$\text{similarity}_r = \cos(\theta_r) = \frac{T_r \cdot R_r}{\|T_r\| \|R_r\|} = \frac{\sum_{i=1}^n T_{ri} \times R_{ri}}{\sqrt{\sum_{i=1}^n (T_{ri})^2} \times \sqrt{\sum_{i=1}^n (R_{ri})^2}} \quad (7.5)$$

The above formula provides the value of similarity between the tag vector and the reference vector of a single resource. This value is the cosine of the angle between the two vectors, which could range from 0 to 1, since the term frequencies only consist of positive values. A value of 1 would mean that both vectors are exactly the same, whereas a 0 would mean they coincide in none of the terms, and so they are completely different. After getting the similarity value between each pair of vectors, we need to get the overall similarity value between users' tags and descriptive data of resources. Accordingly, the similarity between the set of  $n$  reference vectors, and the set of  $n$  tag vectors is computed as the average of similarities between pairs of tag and reference vectors (see Equation 7.6).

$$\text{similarity} = \frac{1}{n} \sum_{r=1}^n \cos(\theta_r) \quad (7.6)$$

This similarity value shows the extent to which the tags provided by the selected set of users resembles the reference descriptive data, i.e., how descriptive

are the tags by those users. The higher is the similarity value, the more descriptive are the tags provided by the users. The closer it is to 0, the more non-descriptive are the tags provided by them. Accordingly, this enables to compare same percents of tag assignments by Categorizers and Describers, and to analyze which of them provide more descriptive tags.

## 7.4 Results

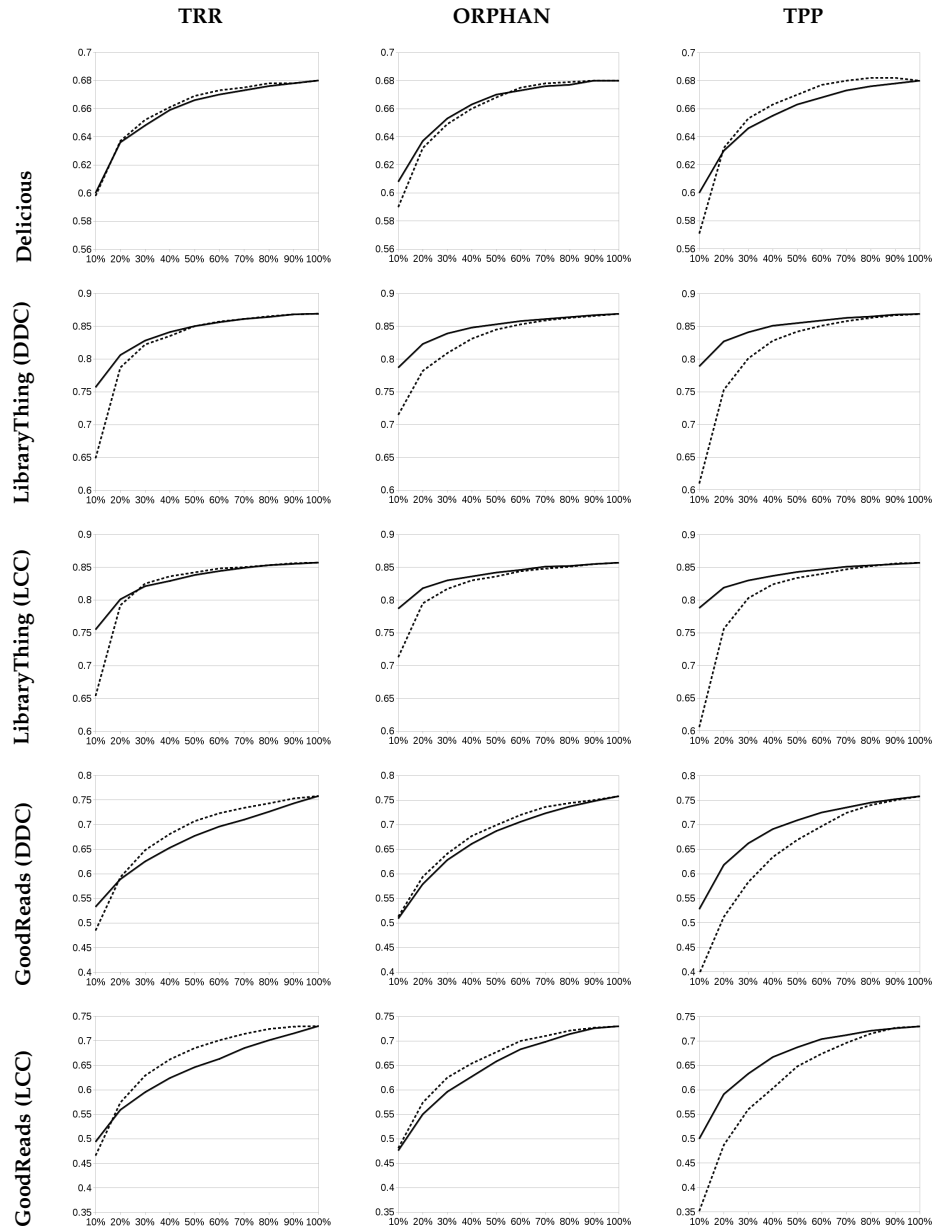
Table 7.3 shows the performance of Categorizers (continuous line) and Describers (dashed line) on the classification task, whereas Table 7.4 on page 121 does the same for the descriptiveness experiments. The results are presented in different graphs organized by datasets in rows –Delicious, LibraryThing and GoodReads–, and by measures in columns –TRR, ORPHAN and TPP. All of them keep the same scale and ranges for X axis, as well as for Y axis within each dataset, so that it enables an easy visual comparison of the results. When analyzing these results, we are especially interested in performance differences between Categorizers and Describers, and studying whether and why such subsets of users perform better for a certain task. Obviously, both Categorizers and Describers always yield the same performance for 100% sets, as we are considering the whole set of users.

### 7.4.1 Categorizers Perform Better on Classification

It stands out that all three measures get positive results for both classification and descriptiveness experiments on LibraryThing. The subsets of Categorizers perform better for classification in all cases for this dataset. This means that all three measures provide a good way to discriminate Categorizers. Among the compared measures, TPP gets the largest gap for classification, whereas TRR does it for descriptiveness.

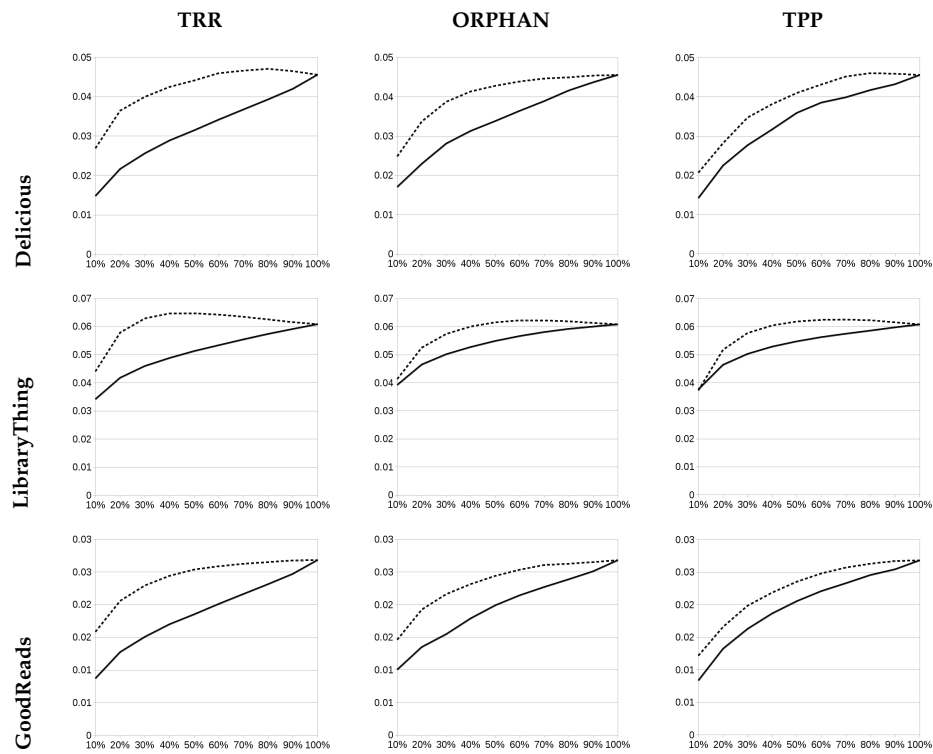
As regards to GoodReads, results are less consistent. TPP yields especially positive results on this dataset. With the other measures, TRR and ORPHAN, Describers outperform Categorizers for classification. However, TRR works well for the 10% subsets, suggesting that it discriminates correctly a subset of extreme Categorizers, but it fails when the subsets upsize. We speculate that the reason for this observation lies in the fact that this social tagging system is suggesting tags to users from their personomy. This encourages users to have a smaller vocabulary, and to reuse their tags frequently. It is quite easier to click on a list of tags than to type them.

In the case of Delicious, it seems that the resource-based system suggestions of its settings make it more difficult to detect Categorizers. On the one hand, TRR and ORPHAN show really slight differences between Categorizers and Describers, so that their discrimination does not seem to be performed appropriately.



**Table 7.3:** Tag-based classification accuracy results for Categorizers (continuous lines) and Describers (dashes lines) on Delicious, LibraryThing and GoodReads. The X axis represents the percents of selected top users, ranging from 10% to 100% with a step size of 10%, either for Categorizers or Describers, whereas Y axis represents the accuracy.





**Table 7.4:** Similarity measures of the descriptiveness of tags on Delicious, LibraryThing and GoodReads. Continuous lines correspond to Categorizers, whereas dashed lines are Describers. The X axis represents the percents of selected top users, ranging from 10% to 100% with a step size of 10%, either for Categorizers or Describers, whereas Y axis represents the degree of similarity (i.e., cosine value) to descriptive data.

However, it works well with TPP for small subsets of users, where Categorizers outscore Describers. This outperformance inverts for larger subsets of users, though.

Analyzing the three datasets altogether, TPP shows the best way to discriminate Categorizers as better contributors to the resource classification task. This is clear for GoodReads and LibraryThing, but it only happens for small subsets of users on Delicious.

#### **7.4.2 Describers Perform Better on Descriptiveness**

The results of the descriptiveness experiments show that Describers are always superior to Categorizers in this regard. All three measures show to be really useful for discriminating Describers among users on social tagging systems, regardless of the settings of the site. Moreover, the measures show a similar behavior on all sites insofar as the outperformance of Describers as compared to Categorizers is fairly similar on the three datasets. However, the large gap of TRR sets it apart from the rest. Thereby, TRR gives rise to the best detection of Describers.

#### **7.4.3 Verbosity vs Diversity**

The three measures we have studied in this work rely on two different features to discriminate user behavior: verbosity and diversity. We can see a better overall performance of the TPP measure for resource classification, and the TRR measure for the descriptiveness task, we believe that: (1) verbosity can be inferred as the optimal feature for discriminating Categorizers, and (2) diversity as the feature that better discriminates Describers. In this context, we believe that Categorizers are thinking of a physical organization of resources, as librarians would do by placing books in shelves, when they annotate resources with tags. For instance, in the specific case of books, a user who thinks of the shelf where they stack their fictional books seems very likely to solely use the tag *fiction*. We could define these shelf-driven users as non-verbose. A user who adds just one tag has probably thought of the perfect tag that places it in the corresponding shelf. On the other hand, users who provide more detailed and diverse annotations rather think of describing the book instead of placing it in a specific shelf. This aspect makes the verbosity feature more powerful than the diversity feature for the detection of Categorizers. Thus, we believe that this is the feature that makes TPP so useful at discriminating Categorizers in search of an accurate resource classification as compared to TRR and ORPHAN, because it only relies on users' verbosity.

#### 7.4.4 Non-descriptive Tags Provide More Accurate Classification

When discriminating user behavior appropriately by using a verbosity-based measure like TPP, we have shown that Categorizers better fit the classification task, whereas Describers provide annotations that further resemble the descriptive data. An interesting deduction from here is that a set of annotations that differs to a greater extent from the descriptive data produces a more accurate classification of the books. From this, we infer that Describers are using more descriptive tags, whereas Categorizers rather use non-descriptive tags. Hence, users who do not think of providing annotations in a similar way to writing reviews rely on non-descriptive tags, yielding a more accurate classification of the books.

### 7.5 Conclusion

In this chapter, we have explored the detection of user behavior on social tagging systems in search of users who rather approach to the resource classification task. To this end, we have explored the measures presented by [Strohmaier et al. \(2010a\)](#), which help us determine whether a user is a Categorizer rather organizing resources, or they are a Describer rather detailing the content of the resources. Specifically, we have studied the application of three different measures –TRR, ORPHAN and TPP–, which rely on two main features: verbosity and diversity of user annotations. By means of choosing different subsets of Categorizers and Describers, we have performed experiments on (1) resource classification, in order to explore whether Categorizers further resemble the classification by experts, and (2) measurement of the descriptiveness of tags, for exploring whether there is a higher similarity between tags by Describers and descriptive data of the resources.

Besides further understanding the existence of users aiming at classification on social tagging systems, i.e., Categorizers, we complemented a previous work by [Körner et al. \(2010a\)](#), where the authors showed that Describers are a good source for inferring semantic relations from folksonomies.

Parts of the research in this chapter have been published in [Zubiaga et al. \(2011b\)](#).

We have answered the following research questions:

#### Research Question 9

*Can we discriminate different user profiles so that we can find a subset of users who provide annotations that better fit a classification scheme?*

We have shown that such type of user, so-called Categorizer, actually exists. Tags assigned by Categorizers provide a more accurate classification of resources than those assigned by another set of users so-called Describers. According to

our experiments, this is mostly true for systems without tag suggestions, i.e., LibraryThing, where the resource classification performed with tags by Categorizers yields clearly better results. When such suggestions exist, the detection of suitable users becomes more difficult, as we have showed happens on GoodReads and Delicious. However, the application of an appropriate measure by considering suitable features can produce a successful selection of users who fit the characteristics of a Categorizer.

**Research Question 10**

*What are the features that identify a user as a good contributor to the resource classification?*

We have analyzed two features that characterize users of social tagging systems: verbosity, and diversity. We have shown that the level of verbosity helps discover Categorizers, who are better suited for the classification task. The vocabulary diversity is useful to find Describers, who tend to annotate using descriptive tags. Moreover, we have shown that users who do not rely on descriptive data provide better classification metadata than those who use descriptive tags.

## Conclusions and Future Research

*“Believe those who are seeking the truth. Doubt those who find it.”*

— Andre Gide

We conclude the thesis in this chapter. Next, we summarize the main contributions to the research field in Section 8.1. We continue by answering the formulated research questions in Section 8.2 on page 127. Finally, in Section 8.3 on page 131 we present an outlook on future directions of the research work in this thesis.

### 8.1 Summary of Contributions

The novel idea of this work lies in the use of social annotations for carrying out a resource classification task. To the best of our knowledge, the first research work performing real classification experiments using social annotations is our first work in the field (Zubiaga et al., 2009d). Prior to that, only Noll and Meinel (2008a) had performed a statistical analysis comparing social tags to a classification performed by experts. Taking into account the lack of work in the field, the work comprised in this thesis sheds new light on the appropriate use and representation of social tags for resource classification. More specifically, the following are the main contributions of this work:

- We have created 3 large-scale social tagging datasets, including classification metadata of the annotated resources. These are among the largest datasets used so far for research and, to the best of our knowledge, the largest used for resource classification experiments. Some of these datasets, along with other smaller datasets we created, have been made publicly available for research purposes<sup>1</sup>. Godoy and Amandi (2010) and Strohmaier

---

<sup>1</sup><http://nlp.uned.es/social-tagging/datasets/>

[et al. \(2010b\)](#), for instance, have used some of our datasets in their recent research works. Even after we created these social tagging datasets, and made publicly available parts of them, little work has been done on creating and releasing more datasets. In [Körner and Strohmaier \(2010\)](#), the authors present a list of publicly available social tagging datasets, among which our datasets are also included. However, the authors set out the problem of the unavailability of more datasets, and encourage researchers to create and release new ones. As far as we know, no additional datasets have been released subsequently including categorization data for tagged resources.

- Our work is the first comparing different representations of resources based on social tags for resource classification. Moreover, it is the first work performing actual classification experiments comparing social tags to other data sources. We have shown that social tags are also useful for classification upon narrower categories in deeper levels of taxonomies. In a previous work, [Noll and Meinel \(2008a\)](#) perform a statistical study concluding that social tags may not be helpful for narrower categories. In contrast to this, we have performed actual classification experiments showing a larger improvement for narrow categorization as compared to other data sources.
- We have analyzed the distributions of social tags in folksonomies, and performed a thorough study on how the settings of each social tagging system affect them, and therefore, a resource classification task. In this regard, we have applied a consolidated weighting scheme, TF-IDF, to the new social data structure given by folksonomies.
- We have shown the existence of a group of users, so-called Categorizers, whose annotations more closely resemble the classification performed by experts than social tags provided by another group of users known as Describers. The approach of differentiating Categorizers from Describers was already tested and verified in earlier works by proving the suitability of the latter for inferring semantic relations from folksonomies. Going further, we have demonstrated the suitability of Categorizers for resource classification.

The use of social annotations for the sake of resource classification tasks was a novel research line in the beginning of this thesis. However, the increasing interest of researchers on user-generated content in social media, and specifically in social tagging systems, has recently brought about more work in the field. Along with this increase, more researchers have shown their interest in the use of social annotations for resource classification tasks, and the number of works in this field has increased. [Godoy and Amandi \(2010\)](#), for instance, perform a tag-based classification study inspired by our earlier work ([Zubiaga et al., 2009d](#)). Furthermore, [Aliakbary et al. \(2009\)](#), [Yin et al. \(2009\)](#), [Xia et al. \(2010\)](#), and [Lu](#)

[et al. \(2010\)](#) have recently presented their research in related matters, making use of social tags as to resource classification.

## 8.2 Answers to Research Questions

At the beginning of this work, we set forth the following problem statement summarizing the main goal of the thesis:

### Problem Statement

*How can the annotations provided by users on social tagging systems be exploited to yield the most accurate resource classification task?*

In order to solve this problem statement, we split it into 10 research questions. Next, we list those research questions along with answers to them:

### Research Question 1

*What kind of SVM classifiers should be used to perform this kind of classification tasks: a native multiclass classifier, or a combination of binary classifiers?*

We have shown the clear superiority of the native multiclass SVM classifiers over the other approaches combining binary classifiers. Our results show that relying on a set of binary classifiers is not a good option when it comes to multiclass taxonomies. Accordingly, native multiclass classifiers, which consider all the classes at the same time and have more knowledge of the whole task, perform much better.

### Research Question 2

*What kind of learning method performs better for this kind of classification tasks: a supervised one, or a semi-supervised one?*

Semi-supervised approaches may perform better when the labeled subset is really small, but supervised approaches, which are computationally less expensive, perform similarly with more labeled documents. Therefore, we have also shown that, unlike binary tasks as shown by [Joachims \(1999\)](#), a supervised approach performs very similar to a semi-supervised approach on these environments. It seems reasonable that predicting the class of uncategorized documents is much more difficult when the number of classes increases, and so the miscategorized documents are harmful for classifier's learning.

Thereby, according to these two conclusions above, we decided to use a supervised multiclass SVM approach.

### Research Question 3

*How do the settings of social tagging systems affect users' annotations and the resulting folksonomies?*

To this end, we have analyzed several features that can be found in different settings of social tagging systems. Among the analyzed features, we have shown the impact of tag suggestions, which considerably alters the resulting folksonomy. In the studied social tagging sites, all of them differ on the settings regarding suggestions:

- **Resource-based suggestions (Delicious):** when the system suggests tags assigned by other users to the resource at the time of bookmarking it, the likelihood of using new tags to further describe such a resource decreases. In this case, users provide less originality and tend to rely on system suggestions.
- **Personomy-based suggestions (GoodReads):** when the system suggests tags previously used by the user, the vocabulary in their personomy tends to be much smaller. However, users do not know how others annotated a resource, and thus they are likely to provide new tags to the resource.
- **Without suggestions (LibraryThing):** when the system does not suggest any tags to the user, the vocabulary in their personomy increases, as well as the diversity of tags in each resource.

#### Research Question 4

*What is the best way of amalgamating users' aggregated annotations on a resource in order to get a single representation for a resource classification task?*

We have shown that it is worthwhile considering all the tags annotated on a resource instead of those in the top that were annotated most. Tags in the top are the most important, and give the main information on the aboutness of resources. However, tags in the tail are helpful to a lesser extent, providing meaningful information and improving the performance of the classifier.

Regarding the weights assigned to those tags when representing a resource, the number of users annotating each tag should be considered in order to get the best results. This is the value that has shown the best results in our experiments. It has outperformed other approaches ignoring weights or considering other data such as the total number of users annotating the resource.

Thereby, the best representation in our experiments is the one that includes all the tags with the values corresponding to the number of users annotating them.

#### Research Question 5

*Despite of the usefulness of social tags for these tasks, is it worthwhile considering their combination with other data sources like the content of the resource as an approach to improve the results even more?*

By means of classifier committees, which combine the predictions by different classifiers, we have shown that tags provide reliable prediction criteria to take



into consideration. SVM classifiers not only predict a category, but also assign a weight to each category based on the given resource. These weights, given in the form of margin values, can be used by other classifiers which rely on different data sources. Adding up weights provided by different classifiers can help predict the correct category when a single classifier fails to categorize the resource appropriately. Weights provided by classifiers relying on social tags are especially useful when combining them with results from other classifiers. Nonetheless, not all data sources are helpful for combination in classifier committees, and the selected data source must be solid enough and provide reliable predictions to outperform the sole use of tags. When data sources are selected appropriately, the performance improvement can be considerable. We have shown that this varies among datasets. For example, with the Delicious dataset, it is important to analyze all three data sources (content, reviews, and tags). However, with the LibraryThing and GoodReads datasets reviews and tags suffice.

#### Research Question 6

*Are social tags also useful and specific enough to classify resources into narrower categories as in deeper levels of hierarchical taxonomies?*

We have analyzed the usefulness of social tags for classification on two different levels of hierarchical taxonomies. Besides broader categories in the top level, we have also explored the classification on narrower categories in the second level. In this regard, social tags have shown to outperform the other data sources on social tagging sites that encourage users to annotate resources (Delicious and LibraryThing). Tags show clear outperformance in these cases, especially on Delicious, where the difference is even more favorable in the second level. This difference is very similar on LibraryThing. Finally, tags from GoodReads do not outperform other data sources at any level because the system does not encourage users to tag books, so that many bookmarks are not annotated.

Our findings provide a different conclusion from that by [Noll and Meinel \(2008a\)](#), where the authors pointed out the hypothesis that social tags were probably useless for deeper levels of taxonomies, and alternative data should be used instead. However, the authors performed just a statistical analysis, and did not confirm the hypothesis with real experiments.

#### Research Question 7

*Can we further consider the distribution of tags across the collection so that we can measure the overall representativity of each tag to represent resources?*

We have analyzed the suitability of IDF-like weighting functions to define the representativity of tags, which consider the distribution of tags through the whole collection of resources. Our experiments have shown that these functions helps improve performance of a resource classification task. However, we have

shown that the settings of the social tagging system have an effect on those distributions. Resource-based tag suggestions have shown to influence the structure of folksonomies greatly. Suggesting tags based on previous annotations of others on the resource causes a very different tag distribution, which in turn, affects the results of the weighting function. When a system enables the resource-based tag suggestions, the use of tag weighting functions performs worse, and combining with other data sources is required to improve performance; this method can even outperform the TF-based approach.

For our classification experiments, we have found that IDF-like weighting functions clearly outperform the TF approach when resource-based tag suggestions are not enabled, i.e., on LibraryThing and GoodReads, both when used on their own, or when combined with other data sources. We found it better to consider just the tag-based approach, without combining them with other data sources, since it provides superior results, which cannot be improved by combining them with other predictions.

#### Research Question 8

*What is the best approach to weigh the representativity of tags in the collection for resource classification?*

Among the studied weighting functions, the one relying on bookmark frequencies has shown to be the best when there are no resource-based tag suggestions. In these cases, IBF performs the best, followed by IRF, and IUF. All of them clearly outperform TF, when both used on their own, and combined with other data sources using classifier committees.

On the other hand, when the social tagging system suggests tags to the user relying on the resource itself, IUF performs better than the others. IUF performs better than IBF and IRF, because of the importance of the ability of users to choose their own tags without relying on suggestions from these systems. Even though IUF does not outperform TF when used on its own, combining it with other data sources produces the best approach. However, it is only slightly better than the committees relying on TF, and any of them can be used to score similar results.

#### Research Question 9

*Can we discriminate different user profiles so that we can find a subset of users who provide annotations that better fit a classification scheme?*

We have shown that such type of user, so-called Categorizer, actually exists. Tags assigned by Categorizers provide a more accurate classification of resources than those assigned by another set of users so-called Describers. According to our experiments, this is mostly true for systems without tag suggestions, i.e., LibraryThing, where the resource classification performed with tags by Categorizers yields clearly better results. When such suggestions exist, the detection of

suitable users becomes more difficult, as we have showed happens on GoodReads and Delicious. However, the application of an appropriate measure by considering suitable features can produce a successful selection of users who fit the characteristics of a Categorizer.

**Research Question 10**

*What are the features that identify a user as a good contributor to the resource classification?*

We have analyzed two features that characterize users of social tagging systems: verbosity, and diversity. We have shown that the level of verbosity helps discover Categorizers, who are better suited for the classification task. The vocabulary diversity is useful to find Describers, who tend to annotate using descriptive tags. Moreover, we have shown that users who do not rely on descriptive data provide better classification metadata than those who use descriptive tags.

## 8.3 Future Directions

The use of social tags for resource classification is still a novel research field with little work done so far. The thesis has shown how social tags can be useful for the resource classification task, and provides analysis to help determine an optimal method to accurately categorize resources based on their social tags. Furthermore, this thesis paves way for future research on the utilization of social tags for resource classification.

Throughout this thesis, we have considered each tag as a different token, regardless of its semantic meaning. In this regard, future work includes analyzing the meaning of each tag trying to discover synonymous words, and relations among them. Either by using natural language processing methods or following ontology-based approaches, it could improve understanding the meaning of each tag and further exploring the knowledge provided by folksonomies.

The three weighting schemes we have used in Chapter 6 on page 95 rely on the classical TF-IDF function designed for text collections. Trying other weighting functions, as well as defining a new one that fits the structure of folksonomies would be also interesting as a future work. This would especially help for systems providing resource-based tag suggestions, like Delicious, where the tested weighting schemes did not perform well.



## Bibliography

- Rabeeh Abbasi, Sergey Chernov, Wolfgang Nejdl, Raluca Paiu, and Steffen Staab. Exploiting flickr tags and groups for finding landmark photos. In *ECIR '09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, pages 654–661, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 978-3-642-00957-0. [100](#)
- Sadegh Aliakbary, Hassan Abolhassani, Hossein Rahmani, and Behrooz Nobakht. Web page classification using social tags. *Computational Science and Engineering, IEEE International Conference on*, 4:588–593, 2009. [43](#), [126](#)
- Ralitsa Angelova, Marek Lipczak, Evangelos Milios, and Pawel Pralat. Characterizing a social bookmarking and tagging network. In *Proc. of the ECAI 2008 Workshop on Mining Social Data (MSoDa)*, pages 21–25. IOS, 2008. [99](#)
- Shenghua Bao, Guirong Xue, Xiaoyuan Wu, Yong Yu, Ben Fei, and Zhong Su. Optimizing web search using social annotations. In *Proceedings of the 16th international conference on World Wide Web, WWW '07*, pages 501–510, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-654-7. [41](#)
- Toine Bogers and Antal van den Bosch. Recommending scientific articles using citeulike. In *Proceedings of the 2008 ACM conference on Recommender systems, RecSys '08*, pages 287–290, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-093-7. [42](#)
- Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, New York, NY, USA, 1992. ACM. ISBN 0-89791-497-X. [36](#)

- Janez Brank, Marko Grobelnik, Natasa Milic-Frayling, and Dunja Mladenic. Interaction of feature selection methods and linear classification models. In *Proceedings of the ICML-02 Workshop on Text Learning*. Forthcoming, 2002. 98
- John S. Breese, David Heckerman, and Carl Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, pages 43–52. Morgan Kaufmann, 1998. 100
- Iván Cantador, Ioannis Konstas, and Joemon M. Jose. Categorising social tags to improve folksonomy-based recommendations. *Web Semant.*, 9:1–15, March 2011. ISSN 1570-8268. doi: <http://dx.doi.org/10.1016/j.websem.2010.10.001>. URL <http://dx.doi.org/10.1016/j.websem.2010.10.001>. 42
- Olivier Chapelle, Mingmin Chi, and Alexander Zien. A continuation method for semi-supervised svms. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 185–192, New York, NY, USA, 2006. ACM. ISBN 1-59593-383-2. 40, 51
- Olivier Chapelle, Vikas Sindhwani, and Sathiya S. Keerthi. Optimization techniques for semi-supervised support vector machines. *Journal of Machine Learning Research*, 9:203–233, June 2008. ISSN 1532-4435. 38
- Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. 36
- Franca Debole and Fabrizio Sebastiani. Supervised term weighting for automated text categorization. In *SAC '03: 2003 ACM symposium on Applied computing*, pages 784–788, New York, NY, USA, 2003. ACM. ISBN 1-58113-624-2. 98
- Jörg Diederich and Tereza Iofciu. Finding communities of practice from user profiles based on folksonomies. In *Proceedings of the 1st International Workshop on Building Technology Enhanced Learning solutions for communities of Practice*, 2006. 100
- Pavel A. Dmitriev, Nadav Eiron, Marcus Fontoura, and Eugene Shekita. Using annotations in enterprise search. In *Proceedings of the 15th international conference on World Wide Web, WWW '06*, pages 811–817, New York, NY, USA, 2006. ACM. ISBN 1-59593-323-9. 41
- Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. Inductive learning algorithms and representations for text categorization. In *CIKM '98: Proceedings of the seventh international conference on Information and knowledge management*, pages 148–155, New York, NY, USA, 1998. ACM. ISBN 1-58113-061-9. 98

- George Forman. Bns feature scaling: an improved representation over tf-idf for svm text classification. In *CIKM*, pages 263–270, 2008. [97](#), [98](#)
- Daniela Godoy and Analía Amandi. Exploiting the social capital of folksonomies for web page classification. In *Software Services for E-World*, volume 341 of *IFIP Advances in Information and Communication Technology*, pages 151–160. Springer, 2010. [43](#), [125](#), [126](#), [163](#), [164](#), [180](#), [181](#)
- Scott Golder and Bernardo A. Huberman. The structure of collaborative tagging systems. *Journal of Information Science*, 32(2), pages 198–208, 2006. [41](#), [63](#), [112](#)
- Manish Gupta, Rui Li, Zhijun Yin, and Jiawei Han. Survey on social tagging techniques. *SIGKDD Explorations*, 12(1):58–72, 2010. [41](#)
- T. Hammond, T. Hannay, B. Lund, and J. Scott. Social bookmarking tools (I). *D-Lib Magazine*, 11(4):1082–9873, 2005a. [113](#)
- Tony Hammond, Timo Hannay, Ben Lund, and Joanna Scott. Social bookmarking tools (I). *D-Lib Magazine*, 11(4):1082–9873, 2005b. [44](#)
- Zelig Harris. Distributional structure. In *Papers in Structural and Transformational Linguistics*, pages 775–794. D. Reidel Publishing Company, Dordrecht, Holland, 1970. [84](#)
- Markus Heckner, Michael Heilemann, and Christian Wolff. Personal information management vs. resource sharing: Towards a model of information behaviour in social tagging systems. In *Third International AAAI Conference on Weblogs and Social Media, ICWSM-09*, pages 42–49, 2009. [112](#), [113](#)
- Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina. Can social bookmarking improve web search? In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 195–206, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-927-9. [22](#), [41](#), [151](#), [168](#)
- Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Information retrieval in folksonomies: Search and ranking. In *The Semantic Web: Research and Applications*, volume 4011 of *LNAI*, pages 411–426. Springer, 2006. [41](#)
- Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13(2), 2002. [39](#)
- Thorsten Joachims. Text categorization with support vector machines: learning with many relevant features. In Claire Nédellec and Céline Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number

- 1398, pages 137–142, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE. [23](#), [52](#), [56](#), [98](#), [151](#), [169](#)
- Thorsten Joachims. Transductive inference for text classification using support vector machines. In *ICML '99: Proceedings of the Sixteenth International Conference on Machine Learning*, pages 200–209, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1-55860-612-2. [37](#), [48](#), [56](#), [127](#), [159](#), [176](#)
- Smola A. J. Kivinen, J. and R. C. Williamson. Online learning with kernels. In S. Becker T. G. Dietterich and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, pages 785–792, Cambridge, MA, 2002. MIT Press. [37](#)
- Christian Körner. Understanding the motivation behind tagging. ACM Student Research Competition - Hypertext'09, 2009. [113](#)
- Christian Körner and Markus Strohmaier. A call for social tagging datasets. *SIG-WEB Newsl.*, pages 2:1–2:6, January 2010. ISSN 1931-1745. [73](#), [126](#)
- Christian Körner, Dominik Benz, Andreas Hotho, Markus Strohmaier, and Gerd Stumme. Stop Thinking, Start Tagging: Tag Semantics Emerge from Collaborative Verbosity. In *International World Wide Web Conference*, pages 521–530, 2010a. [44](#), [113](#), [123](#)
- Christian Körner, Roman Kern, Hans-Peter Grahsl, and Markus Strohmaier. Of Categorizers and Describers: An Evaluation of Quantitative Measures for Tagging Motivation. In *Conference on Hypertext and Hypermedia*, pages 157–166, 2010b. [44](#), [112](#), [113](#)
- Man Lan, Chew-Lim Tan, Hwee-Boon Low, and Sam-Yuan Sung. A comprehensive comparative study on term weighting schemes for text categorization with support vector machines. In *WWW '05: 14th international conference on World Wide Web*, pages 1032–1033, New York, NY, USA, 2005. ACM. ISBN 1-59593-051-5. [97](#), [98](#)
- Xin Li, Lei Guo, and Yihong Eric Zhao. Tag-based social interest discovery. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 675–684, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-085-2. [41](#), [99](#)
- Huizhi Liang, Yue Xu, Yuefeng Li, Richi Nayak, and Xiaohui Tao. Connecting users and items with weighted tags for personalized item recommendations. In *HT '10: Proceedings of the 21st ACM conference on Hypertext and hypermedia*, pages 51–60, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0041-4. [99](#), [100](#)



- Caimei Lu, Xin Chen, and E. K. Park. Exploit the tripartite network of social tagging for web clustering. In *Proceeding of the 18th ACM conference on Information and knowledge management, CIKM '09*, pages 1545–1548, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-512-3. [44](#)
- Caimei Lu, Jung-ran Park, and Xiaohua Hu. User tags versus expert-assigned subject terms: A comparison of librarything tags and library of congress subject headings. *Journal of Information Science*, 36(6):763–779, 2010. [44](#), [126](#)
- C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, New York, NY, USA, 2006a. ACM. ISBN 1-59593-417-0. [44](#)
- Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *Proceedings of the seventeenth conference on Hypertext and hypermedia, HYPERTEXT '06*, pages 31–40, New York, NY, USA, 2006b. ACM. ISBN 1-59593-417-0. [112](#), [113](#)
- Michael G. Noll and Christoph Meinel. Authors vs. readers: A comparative study of document metadata and content in the www. In *Proceedings of the 2007 ACM symposium on Document engineering*, pages 177–186, Winnipeg, Manitoba, Canada, 2007. ACM. ISBN 978-1-59593-776-6. [43](#)
- Michael G. Noll and Christoph Meinel. Exploring social annotations for web document classification. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 2315–2320, Fortaleza, Ceara, Brazil, 2008a. ACM. ISBN 978-1-59593-753-7. [42](#), [77](#), [78](#), [93](#), [125](#), [126](#), [129](#), [161](#), [163](#), [178](#), [180](#)
- Michael G. Noll and Christoph Meinel. The metadata triumvirate: Social annotations, anchor texts and search queries. In *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on*, volume 1, pages 640–647, 2008b. [43](#), [60](#)
- Oded Nov, Mor Naaman, and Chen Ye. Motivational, structural and tenure factors that impact online community photo sharing. In *Proceedings of ICWSM 2009, the Third International Conference on Weblogs and Social Media*, pages 138–145, 2009. [112](#)
- Martin F. Porter. An Algorithm for Suffix Stripping. *Program*, 14(3):130–137, 1980. [84](#)
- Heng-Nian Qi, Jian-Gang Yang, Yi-Wen Zhong, and Chao Deng. Multi-class svm based remote sensing image classification and its semi-supervised improvement scheme. In *Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on*, volume 5, pages 3146–3151, August 2004. [40](#), [50](#)

- Xiaoguang Qi and Brian D. Davison. Web page classification: Features and algorithms. *ACM Comput. Surv.*, 41(2):1–31, 2009. ISSN 0360-0300. [21](#), [150](#), [168](#)
- Daniel Ramage, Paul Heymann, Christopher D. Manning, and Hector Garcia-Molina. Clustering the tagged web. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 54–63, Barcelona, Spain, 2009. ACM. ISBN 978-1-60558-390-7. [43](#), [78](#), [99](#)
- G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18:613–620, November 1975. ISSN 0001-0782. [97](#)
- Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24:513–523, August 1988. ISSN 0306-4573. [52](#), [97](#)
- Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola, editors. *Advances in kernel methods: support vector learning*. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-19416-3. [37](#)
- Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, 2002. ISSN 0360-0300. [34](#)
- Shilad Sen, Shyong K. Lam, Al Mamunur Rashid, Dan Cosley, Dan Frankowski, Jeremy Osterhouse, F. Maxwell Harper, and John Riedl. tagging, communities, vocabulary, evolution. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work, CSCW '06*, pages 181–190, New York, NY, USA, 2006. ACM. ISBN 1-59593-249-6. [112](#)
- Andriy Shepitsen, Jonathan Gemmell, Bamshad Mobasher, and Robin Burke. Personalized recommendation in social tagging systems using hierarchical clustering. In *RecSys '08: 2008 ACM conference on Recommender Systems*, pages 259–266, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-093-7. [41](#), [99](#)
- Mark Sinka and David Corne. *A Large Benchmark Dataset for Web Document Clustering*, pages 881–890. 2002. [51](#)
- Gene Smith. *Tagging: people-powered metadata for the social web*. New Riders, Berkeley, California, 2008. ISBN 978-0-321-52917-6. [24](#), [41](#), [42](#), [153](#), [170](#)
- Pascal Soucy. Beyond tfidf weighting for text categorization in the vector space model. In *IJCAI 2005: Proc. of the 19th International Joint Conference on Artificial Intelligence*, pages 1130–1135, 2005. [98](#)
- M. Strohmaier, C. Körner, and R. Kern. Why do users tag? detecting users' motivation for tagging in social tagging systems. In *International AAAI Conference*

- on Weblogs and Social Media (ICWSM2010)*, Washington, DC, USA, 2010a. [112](#), [123](#)
- Markus Strohmaier, Christoph Trattner, Denis Helic, and Keith Andrews. Network-theoretic potentials and limitations of tag clouds as a tool for social navigation. *Journal of Universal Computer Science*, 2010b. [125](#), [163](#), [180](#)
- Bing-Yu Sun, De-Shuang Huang, Lin Guo, and Zhong-Qiu Zhao. Support vector machine committee for classification. In *Advances in Neural Networks - ISNN 2004*, pages 648–653, 2004. [86](#)
- Chade-Meng Tan, Yuan-Fang Wang, and Chan-Do Lee. The use of bigrams to enhance text categorization. *Inf. Process. Manage.*, 38(4):529–546, 2002. ISSN 0306-4573. [52](#)
- Jason Weston and Chris Watkins. Multi-class support vector machines. In *Proceedings of the 1999 European Symposium on Artificial Neural Networks*, 1999. [38](#), [39](#)
- Jing Xia, Kunmei Wen, Ruixuan Li, and Xiwu Gu. Optimizing academic conference classification using social tags. *Computational Science and Engineering, IEEE International Conference on*, 0:289–294, 2010. [44](#), [126](#)
- Linli Xu and Dale Schuurmans. Unsupervised and semi-supervised multi-class support vector machines. *AAAI*, 2005. [40](#), [50](#)
- Zenglin Xu, Rong Jin, Jianke Zhu, Irwin King, and Michael Lyu. Efficient convex relaxation for transductive support vector machine. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, pages 1641–1648, Cambridge, MA, 2008. MIT Press. [37](#)
- Yasutoshi Yajima and Tien-Fang Kuo. Optimization approaches for semi-supervised multiclass classification. In *ICDMW '06: Proceedings of the Sixth IEEE International Conference on Data Mining - Workshops*, pages 863–867, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 0-7695-2702-7. [40](#), [49](#), [55](#)
- Yiming Yang and Xin Liu. A re-examination of text categorization methods. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49, New York, NY, USA, 1999. ACM. ISBN 1-58113-096-1. [98](#)
- Zhijun Yin, Rui Li, Qiaozhu Mei, and Jiawei Han. Exploring social tagging graph for web object classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '09*, pages 957–966, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-495-9. [44](#), [126](#)

- Yin Zhang, Kening Gao, Bin Zhang, Jinhua Guo, Feihang Gao, and Pengwei Guo. Clustering blog posts using tags and relations in the blogosphere. In *Proceedings of the 2009 First IEEE International Conference on Information Science and Engineering, ICISE '09*, pages 817–820, Washington, DC, USA, 2009. IEEE Computer Society. ISBN 978-0-7695-3887-7. [44](#)
- Arkaitz Zubiaga. Enhancing Navigation on Wikipedia with Social Tags. In *Wikimania 2009: 4th Annual Conference of the Wikimedia Community*, August 2009. [42](#)
- Arkaitz Zubiaga, Víctor Fresno, and Raquel Martínez. Comparativa de Aproximaciones a SVM Semisupervisado Multiclase para Clasificación de Páginas Web. 42:63–70, 2009a. [56](#)
- Arkaitz Zubiaga, Víctor Fresno, and Raquel Martínez. Is unlabeled data suitable for multiclass svm-based web page classification? In *SemiSupLearn '09: Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 28–36, Morristown, NJ, USA, 2009b. Association for Computational Linguistics. ISBN 978-1-932432-38-1. [56](#)
- Arkaitz Zubiaga, Raquel Martínez, and Víctor Fresno. Clasificación de páginas web con anotaciones sociales. In *Proceedings of SEPLN 2009: The 25th edition of the Annual Conference of the Spanish Society for Natural Language Processing*, 2009c. [92](#)
- Arkaitz Zubiaga, Raquel Martínez, and Víctor Fresno. Getting the most out of social annotations for web page classification. In *Proceedings of the 9th ACM symposium on Document engineering, DocEng '09*, pages 74–83, New York, NY, USA, 2009d. ACM. ISBN 978-1-60558-575-8. [43](#), [92](#), [125](#), [126](#), [163](#), [164](#), [180](#), [181](#)
- Arkaitz Zubiaga, Víctor Fresno, and Raquel Martínez. Exploiting social annotations for resource classification. In I-Hsien Ting, Tzung-Pei Hong, and Leon S.L. Wang, editors, *Social Network Mining, Analysis and Research Trends: Techniques and Applications*, pages 447–497. IGI Global, 2011a. [92](#)
- Arkaitz Zubiaga, Christian Körner, and Markus Strohmaier. Tags vs Shelves: From Social Tagging to Social Classification. In *HT 2011: Proceedings of the 22st ACM Conference on Hypertext and Hypermedia*, 2011b. [123](#)

## Publications

### Peer-Reviewed Conferences

- Arkaitz Zubiaga, Christian Körner, Markus Strohmaier. 2011. *Tags vs Shelves: From Social Tagging to Social Classification*. In Proceedings of Hypertext 2011, the 22nd ACM Conference on Hypertext and Hypermedia, Eindhoven, Netherlands. (acceptance rate: 35/104, 34%)
- Arkaitz Zubiaga, Raquel Martínez, Víctor Fresno. 2009. *Getting the Most Out of Social Annotations for Web Page Classification*. In Proceedings of DocEng 2009, the 9th ACM Symposium on Document Engineering, pp. 74-83, Munich, Germany. (acceptance rate: 16/54, 29.6%)
- Arkaitz Zubiaga, Raquel Martínez, Víctor Fresno. 2009. *Clasificación de Páginas Web con Anotaciones Sociales*. In Proceedings of SEPLN 2009, XXV edición del Congreso Anual de la Sociedad Española para el Procesamiento del Lenguaje Natural, pp. 225-233, Donostia-San Sebastián. (acceptance rate: 36/72, 50%)
- Arkaitz Zubiaga, Alberto P. García-Plaza, Víctor Fresno, Raquel Martínez. 2009. *Content-based Clustering for Tag Cloud Visualization*. In Proceedings of ASONAM 2009, International Conference on Advances in Social Networks Analysis and Mining, pp. 316-319, Athens, Greece.

### Workshops

- Arkaitz Zubiaga, Víctor Fresno, Raquel Martínez. 2009. *Is Unlabeled Data Suitable for Multiclass SVM-based Web Page Classification?*. In Proceedings of

the NAACL-HLT 2009 Workshop on Semi-supervised Learning for Natural Language Processing, pp. 28-36, Boulder, CO, United States.

## Book Chapters

- Arkaitz Zubiaga, Víctor Fresno, Raquel Martínez. 2011. *Exploiting Social Annotations for Resource Classification*. Social Network Mining, Analysis and Research Trends: Techniques and Applications. IGI Global.

## Journals

- Arkaitz Zubiaga, Víctor Fresno, Raquel Martínez. 2009. Comparativa de Aproximaciones a SVM Semisupervisado Multiclase para Clasificación de Páginas Web. SEPLN, Sociedad Española para el Procesamiento del Lenguaje Natural, vol. 42, pp. 63-70.

## Others

- Arkaitz Zubiaga. 2009. *Enhancing Navigation on Wikipedia with Social Tags*. Wikimania 2009, Buenos Aires, Argentina.
- Arkaitz Zubiaga, Alberto P. García-Plaza, Víctor Fresno, Raquel Martínez. 2009. Etiketa-lainoen Ikuskera Hobetzeko Multzokatzea. Informatikari Euskaldunen Bilkura '09, Donostia-San Sebastián.



## Additional Results

In [Chapter 5 on page 75](#) we explored different representations of social tags in order to evaluate which of them performs better on a resource classification task. Among the approaches, we compared using all the tags annotated on each resource, and choosing just those in the top. For the latter, we focused on the top 10 tags, just to evaluate whether tags in the tail were harmful for this purpose. However, we did not show whether a selection of top 5 or 15 of tags could be a better choice. In [Table A.1 on the next page](#) we show the results of using different tops of tags for the FTA-based representation on the top level of the taxonomies. The results confirm that relying on all the tags performs the best, and that the selection of 5, 10 or 15 tags in the top has no impact in this regard. Going further, it also confirms that tags in the tail are far less useful, because the improvement is much smaller when low-ranked tags are included.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Delicious	.445	.542	.583	.602	.620	.633	.639	.648	.652	.654	.659	.661	.665	.665	.666
LibraryThing (DDC)	.805	.849	.855	.857	.858	.860	.861	.862	.864	.864	.865	.864	.864	.865	.865
LibraryThing (LCC)	.778	.833	.844	.848	.852	.854	.855	.857	.857	.858	.859	.858	.858	.858	.859
GoodReads (DDC)	.660	.714	.725	.731	.730	.730	.730	.730	.728	.730	.730	.729	.730	.730	.730
GoodReads (LCC)	.619	.687	.707	.709	.709	.710	.711	.709	.711	.712	.712	.710	.711	.713	.711
	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Delicious	.668	.670	.670	.673	.672	.673	.673	.674	.675	.675	.675	.677	.677	.677	.677
LibraryThing (DDC)	.866	.866	.866	.867	.866	.867	.866	.866	.867	.867	.866	.866	.866	.867	.866
LibraryThing (LCC)	.859	.860	.860	.860	.860	.859	.860	.859	.860	.860	.860	.860	.860	.860	.860
GoodReads (DDC)	.730	.729	.729	.730	.729	.730	.731	.730	.729	.728	.729	.730	.730	.729	.730
GoodReads (LCC)	.711	.711	.715	.713	.711	.710	.713	.709	.712	.711	.713	.712	.713	.712	.713
	40	50	60	70	80	90	100	150	200	250	300	350	400	450	500
Delicious	.678	.679	.678	.678	.679	.679	.679	.679	.679	.679	.680	.680	.679	.679	.679
LibraryThing (DDC)	.867	.867	.867	.867	.868	.868	.868	.867	.868	.868	.868	.868	.868	.867	.868
LibraryThing (LCC)	.860	.861	.861	.860	.861	.861	.861	.861	.861	.861	.860	.861	.861	.861	.861
GoodReads (DDC)	.729	.729	.728	.729	.729	.728	.728	.729	.729	.729	.730	.730	.730	.730	.728
GoodReads (LCC)	.713	.713	.713	.709	.710	.709	.714	.711	.710	.712	.713	.713	.711	.713	.710

Table A.1: Accuracy results of tag-based classification relying with different number of tags in the top.





## Key Terms and Definitions

Next, we list and provide the definitions for some of the most relevant terms related to this thesis, which help to better understand social tagging systems:

**Tagging** Tagging is an open way to assign tags or keywords to resources or items (e.g., web pages, movies or books), in order to describe them. This enables the later retrieval of the resources in an easier way, using tags as resource metadata. As opposed to a classical taxonomy-based categorization system, they are usually non-hierarchical, and the vocabulary is open, so it tends to grow indefinitely. For instance, a user could tag this thesis as social-tagging, research and thesis, whereas another user could use web2.0, social-bookmarking and tagging tags to annotate it.

**Social tagging** A tagging system becomes social when its tag annotations are publicly visible, and profitable for anyone. The fact of a tagging system being social implies that a user could take advantage of tags defined by others to retrieve a resource.

**Social bookmarking** Delicious, StumbleUpon and Diigo, amongst others, are known as social bookmarking sites. They provide a social means to save web pages (or other online resources like images or videos) as bookmarks, in order to retrieve them later on. In contrast to saving bookmarks in user's local browser, posting them to social bookmarking sites allows the community to discover others' links and, besides, to access the bookmarks from any computer to the user itself. In these systems, bookmarks represent references to web resources, and do not attach a copy of them, but just a link. Note that social bookmarking sites do not always rely on social tags to organize resources, e.g., Reddit is a social bookmarking approach to add comments on web pages instead of tags. However, the use of social tags in social bookmarking systems is a common approach.

**Social cataloging** They are quite similar to social bookmarking sites in that resources are socially shared but, in this case, offline resources like music, books or movies are saved. For instance, LibraryThing allows to save the books you like, Hulu does it for movies and TV series, and Last.fm for music-related resources. As in social bookmarking sites, tags are the most common way to annotate resources in social cataloging sites.

**Folksonomy** As a result of a community tagging resources, the collection of tags defined by them creates a tag-based organization, so-called folksonomy. A folksonomy is also known as a community-based taxonomy, where the classification scheme is plain, there are no predefined tags, and therefore users can freely choose new words as tags. A folksonomy is basically known as weighted set of tags, and may refer to a whole collection/site, a resource or a user. A summary of a folksonomy is usually presented in the form of a tag cloud.

**Personomy** Personomy is a neologism created from the term folksonomy, and it refers to the weighted set of tags of a single user/person. It summarizes the topics a user tags about.

**Simple tagging** users describe their own resources or items, such as photos on Flickr, news on Digg or videos on Youtube, but nobody else tags another user's resources. Usually, the author of the resource is who tags it. This means no more than one user tags an item. In many cases, like in Flickr and Youtube, simple tagging systems include an attachment to the resource, and not just a reference to it.

**Collaborative tagging** many users tag the same item, and every person can tag it with their own tags in their own vocabulary. The collection of tags assigned by a single user creates a smaller folksonomy, also known as personomy. As a result, several users tend to post the same item. For instance, CiteULike, LibraryThing and Delicious are based on collaborative tagging, where each resource (papers, books and URLs, respectively) could be annotated by all the users who considered it interesting.



## List of Acronyms

This is a list of acronyms used in this thesis:

**API** Application Programming Interface

**DDC** Dewey Decimal Classification

**FTA** Full Tagging Activity

**HTML** HyperText Markup Language

**IBF** Inverse Bookmark Frequency

**IDF** Inverse Document Frequency

**IRF** Inverse Resource Frequency

**IUF** Inverse User Frequency

**LCC** Library of Congress Classification

**ODP** Open Directory Project

**ORPHAN** Orphan Ratio

**TPP** Tags Per Post

**TRR** Tag Resource Ratio

**URL** Uniform Resource Locator

**SVM** Support Vector Machines

**S<sup>3</sup>VM** Semi-Supervised Support Vector Machines

**TF** Term Frequency

**VSM** Vector Space Model





## Resumen (Spanish Summary)

*“El experimentador que no sabe lo que está buscando no comprenderá lo que encuentra.”*

— Claude Bernard

### Utilización de Folksonomías para Clasificación de Recursos

En esta tesis abordamos el problema de la clasificación automática de recursos, una tarea cada vez más importante en nuestra vida diaria. El catalogado de libros o la organización de vídeos, entre otros, representan algunos ejemplos de actividades para las que un proceso automático de clasificación resulta cada vez más necesario e importante en nuestro día a día. En esta tesis aprovechamos la información contenida en las anotaciones que realizan los usuarios de sistemas de etiquetado social, en los cuales se recogen metadatos que detallan el contenido de diferentes tipos de recursos, para mejorar la clasificación. Hasta el momento, son pocos los trabajos que han explotado estos metadatos con este fin, y los pocos que lo han hecho se han limitado a realizar análisis estadísticos. En esta tesis exploramos las características de estos sistemas de etiquetado social y de los usuarios involucrados en ellos, así como de las anotaciones que aportan, siempre con el fin de sacar el máximo partido a estas grandes colecciones, obteniendo así el mayor rendimiento posible para un clasificador automático de recursos.

## D.1 Motivación

Organizar recursos dentro de categorías supone una tarea muy común en nuestro día a día. Tener recursos asignados a categorías predefinidas siempre ayuda a mejorar posteriores accesos a la información contenida en ellos, ya que este acceso puede limitarse entonces a un conjunto reducido de categoría(s) deseada(s). Por ejemplo, los bibliotecarios suelen catalogar los libros por temas, de forma que quedan organizados por intereses similares. Las bases de datos de películas, los catálogos de música y los sistemas de ficheros, entre otros, suelen estar organizados también por categorías, de forma que se facilita su acceso futuro. Asimismo, la clasificación de páginas web resulta una tarea de especial interés a la hora de mejorar los resultados provistos por los motores de búsqueda, ya que ayudan a reducir el ámbito de esta búsqueda a la categoría deseada por el usuario. Directorios web como Yahoo! Directory y Open Directory Project organizan páginas web en categorías, ofreciendo una alternativa o complemento a la búsqueda por palabra(s) clave(s).

El problema entonces está en lo costosa y cara que resulta la clasificación manual de estos recursos cuando la colección es grande. Por ejemplo, The Library of Congress de Estados Unidos informó en 2002 de que el coste medio de catalogación de cada registro bibliográfico por profesionales fue de 94,58 dólares<sup>1</sup>. Catalogar 291.749 registros, como hicieron en aquel año, les llegó a costar unos 27 millones y medio de dólares. Dado lo cara que resulta la categorización manual, la utilización de clasificadores automáticos puede ser una buena alternativa para reducir su coste, y asimismo mantener los catálogos al día con un esfuerzo humano menor.

Hasta el momento, la mayoría de los clasificadores automáticos se han centrado en el contenido de los recursos a la hora de representarlos, sobre todo en tareas de clasificación de páginas web (Qi and Davison (2009)). No obstante, la falta de datos representativos en el contenido de muchos de ellos hace que se complique esta tarea. Además, puede resultar muy complicado obtener suficientes datos sobre tipos de recursos como libros o películas, para los cuales puede ser más complicado representar el contenido o, incluso, puede que el contenido no esté disponible en una forma que pueda ser procesado.

Como solución a este problema, los sistemas de etiquetado social proveen una forma sencilla y barata de obtener metadatos sobre recursos. Sistemas como Delicious<sup>2</sup>, LibraryThing<sup>3</sup> y GoodReads<sup>4</sup> recopilan anotaciones de usuarios en forma de etiquetas para grandes colecciones de recursos. Estas etiquetas provistas

---

<sup>1</sup><http://www.loc.gov/loc/lcib/0302/collections.html>

<sup>2</sup><http://delicious.com>

<sup>3</sup><http://www.librarything.com>

<sup>4</sup><http://www.goodreads.com>

por usuarios dan lugar a datos significativos que describen el contenido de los recursos (Heymann et al., 2008).

Por medio de estas etiquetas, los usuarios proveen una especie de organización propia de los recursos. Estas etiquetas se comparten de forma social con la comunidad, y gracias a que un gran número de usuarios contribuye en estos sistemas, son numerosas las anotaciones que se acumulan sobre cada recurso. Por lo tanto, esa acumulación hace que cada una de las anotaciones sea más útil. Así, la acumulación de usuarios en una comunidad activa genera un gran número de marcadores, etiquetas, y por tanto, recursos anotados.

*“Cada una de las categorizaciones individuales vale menos que la categorización de un profesional. Pero hay muchas, muchas de aquéllas.”*, Joshua Schachter, fundador de Delicious, en la cumbre FOWA 2006 FOWA en Londres (Inglaterra)<sup>5</sup>.

Los sistemas de etiquetado social representan un medio para guardar, organizar y buscar recursos, todo ello por medio de la anotación con etiquetas escogidas por el usuario. Como hipótesis principal de este trabajo, creemos que estas grandes colecciones de anotaciones pueden mejorar de forma considerable una tarea de clasificación de recursos. Dicho de otro modo, las anotaciones provistas por usuarios podrían llegar a ser muy útiles como una fuente de datos que aporta información significativa que podría ayudar a inferir la categoría de los recursos.

Dado que un gran número de usuarios provee sus propias anotaciones sobre cada recurso, nuestro objetivo entonces se centra en descubrir la manera de amalgamar esas aportaciones en busca de una organización que se parezca a la categorización realizada por profesionales. En este contexto, donde los usuarios aportan grandes cantidades de metadatos, nuestro reto se centra en sacar el máximo partido de ellos con el fin de mejorar el rendimiento de la tarea de clasificación de recursos.

*“Estamos en una época en la que los datos son baratos, pero sacar partido de ellos no lo es”*, Danah Boyd, Investigadora sobre Social Media en Microsoft Research New England, en el congreso WWW2010 en Raleigh, Carolina del Norte, Estados Unidos<sup>6</sup>.

### D.1.1 Clasificación de Recursos

La clasificación de recursos se puede definir como la tarea consistente en la organización de recursos dentro de un conjunto de categorías predefinidas. En este trabajo utilizamos las Máquinas de Vectores de Soporte (SVM, Joachims (1998)),

<sup>5</sup><http://simonwillison.net/2006/Feb/8/summit/>

<sup>6</sup><http://www.danah.org/papers/talks/2010/WWW2010.html>

un método vanguardista para clasificación que ha destacado por sus buenos resultados desde finales de los años 90. Este algoritmo de clasificación se basa en el análisis de un conjunto de instancias previamente categorizadas, con lo que se alimenta el clasificador para que adquiera el conocimiento necesario para poder clasificar posteriormente nuevos recursos.

Un problema de clasificación de recursos puede definirse a partir de diferentes características. Por una parte, en lo que se refiere al método de aprendizaje, puede ser *supervisado*, donde todo el conjunto de entrenamiento está previamente categorizado, o *semisupervisado*, donde también se aprovechan instancias sin información de categoría durante la fase de aprendizaje. Por otra parte, considerando el número de clases, la clasificación puede ser *binaria*, cuando sólo hay dos categorías que pueden ser asignadas a cada recurso, o *multiclase*, cuando hay tres o más categorías. El primer caso se utiliza habitualmente para sistemas de filtrado, mientras que el segundo suele ser frecuente en el caso de taxonomías mayores, como en el caso de la clasificación temática de recursos.

Para clasificación temática sobre grandes colecciones de recursos, como páginas web en la Web o libros en bibliotecas, las taxonomías suelen estar definidas por más de dos categorías, y el subconjunto de recursos previamente categorizado suele ser muy pequeño. De esta manera, creemos que se debería considerar y analizar la aplicación de técnicas semisupervisadas y multiclase para este tipo de tareas.

Por ello, en esta tesis proponemos inicialmente el análisis de varias técnicas de clasificación que utilizan SVM, con el fin de analizar su adecuación a estas tareas. Estas técnicas incluyen diferentes aproximaciones a la resolución de tareas multiclase, así como algoritmos supervisados y semisupervisados.

### D.1.2 Anotaciones Sociales

Los sistemas de etiquetado social permiten a sus usuarios guardar y anotar sus recursos favoritos (como por ejemplo páginas web, películas, libros, fotos o música), compartiéndolos a su vez con la comunidad. Los usuarios proveen estas anotaciones normalmente en forma de etiquetas. Se conoce como etiquetado a la forma abierta de asignar etiquetas o palabras clave a recursos, de manera que se pueden describir y organizar. Esto posibilita la posterior recuperación de los recursos de forma más sencilla, aprovechando las etiquetas como metadatos que los describen. Normalmente, no hay etiquetas predefinidas, y por lo tanto los usuarios pueden escoger libremente las palabras que deseen como etiquetas.

*“El etiquetado es principalmente una interfaz de usuario - una manera para que la gente recuerde cosas, en qué estaban pensando en el momento en el que lo guardaron. Bastante útil para recordar, bueno para el descubrimiento, terrible para la distribución (donde los que lo publican añaden tantas eti-*



*quetas como pueden para incluirlo en el mayor número posible de cajas).”,*  
Joshua Schachter, fundador de Delicious, en la cumbre FOWA 2006  
FOWA en Londres (Inglaterra)<sup>7</sup>.

Mediante este proceso se genera una estructura de etiquetas conocida como folksonomía, es decir, una organización de recursos dirigida por usuarios. Folksonomía es una contracción de las palabras *folk* (gente), *taxis* (clasificación) y *nomos* (gestión). Es conocida también como una taxonomía basada en los usuarios, en la cual la estructura no es jerárquica, al contrario que una clasificación taxonómica básica. Por lo tanto, una folksonomía tiene cierta relación con las taxonomías generadas por expertos, en cuanto a que los recursos se organizan igualmente en grupos.

Se dice que estas anotaciones pertenecen a un entorno social cuando están accesibles y utilizables para cualquier usuario. Esta característica posibilita la búsqueda de recursos aprovechando las anotaciones aportadas por otros. A su vez, es uno de los motivos que anima a los usuarios a contribuir.

No obstante, no todas las anotaciones se comparten de la misma manera. El propio sistema de etiquetado social puede definir algunas restricciones a este respecto, principalmente estableciendo quién tiene permiso para anotar cada recurso. En este sentido, se pueden distinguir dos tipos de sistemas (Smith, 2008):

- **Sistemas de etiquetado simple:** los usuarios pueden describir sus propios recursos, como es el caso del etiquetado de fotos en Flickr<sup>8</sup>, noticias en Digg<sup>9</sup> o vídeos en Youtube<sup>10</sup>, pero nadie anota los recursos de otros. Generalmente, el autor del recurso es quien lo anota. Esto significa que no más de un usuario puede etiquetar cada recurso. Más formalmente, en un sistema de etiquetado simple hay un conjunto de usuarios ( $U$ ) que anota unos recursos ( $R$ ) con unas etiquetas ( $T$ ). Cada usuario  $u_i \in U$  puede guardar un recurso  $r_j \in R$  con un conjunto de etiquetas  $T_j = \{t_{j1}, \dots, t_{jp}\}$ , con un número  $p$  variable de etiquetas. El conjunto de etiquetas asignado a  $r_j$  seguirá estando limitado a  $T_j$ , ya que nadie más lo podrá anotar.
- **Sistemas de etiquetado colaborativo:** muchos usuarios pueden anotar cada recurso, y todos ellos pueden etiquetarlo con su propio vocabulario. El conjunto de etiquetas asignado por un usuario genera una folksonomía a menor escala, conocida como personomía. Como resultado, varios usuarios tienden a anotar el mismo recurso. Por ejemplo, CiteULike.org, LibraryThing.com y Delicious se basan en anotaciones colaborativas, donde cada recurso (artículos, libros y URLs, respectivamente) puede ser anotado y

<sup>7</sup><http://simonwillison.net/2006/Feb/8/summit/>

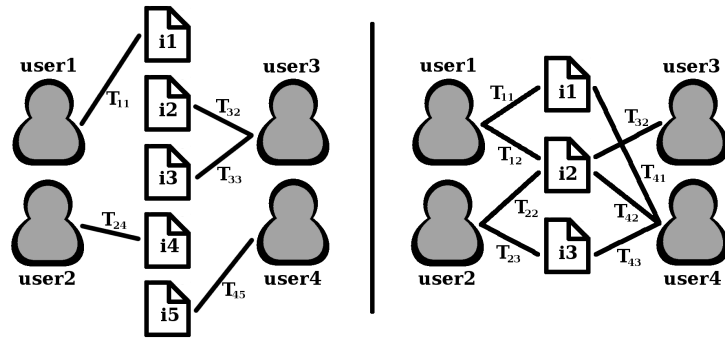
<sup>8</sup><http://www.flickr.com>

<sup>9</sup><http://digg.com>

<sup>10</sup><http://www.youtube.com>

etiquetado por todos aquellos usuarios que lo consideren interesante. Por tanto, los sistemas de etiquetado colaborativo son algo más complejos, ya que hay un conjunto de usuarios ( $U$ ) que guarda sus marcadores ( $B$ ) sobre unos recursos ( $R$ ) anotándolos con unas etiquetas ( $T$ ). Cada usuario  $u_i \in U$  puede guardar un marcador  $b_{ij} \in B$  de un recurso  $r_j \in R$  con un conjunto de etiquetas  $T_{ij} = \{t_{ij1}, \dots, t_{ijp}\}$ , con un número  $p$  variable de etiquetas. Después de que  $k$  usuarios guardan  $r_j$ , se describe como un conjunto pesado de etiquetas  $T_j = \{w_{j1}t_{j1}, \dots, w_{jn}t_{jn}\}$ , donde  $w_{j1}, \dots, w_{jn} \leq k$  representan el número de asignaciones de cada etiqueta. Por lo tanto, cada marcador está compuesto por la tripleta de un usuario, un recurso y un conjunto de etiquetas:  $b_{ij} : u_i \times r_j \times T_{ij}$ . Así, cada usuario guarda marcadores de diferentes recursos, y cada recurso tiene marcadores correspondientes a diferentes usuarios. El resultado de acumular etiquetas contenidas en los marcadores de un usuario se conoce como la personomía de ese usuario:  $T_i = \{w_{i1}t_{i1}, \dots, w_{im}t_{im}\}$ , donde  $m$  es el número de etiquetas diferentes en la personomía del usuario.

La Figura D.1 muestra un ejemplo comparativo de ambos tipos de sistemas.



**Figura D.1:** Comparación de anotaciones provistas por usuarios en sistemas de etiquetado simple y colaborativo.

En esta tesis nos centramos en sistemas de etiquetado colaborativo. Generalmente, las etiquetas asociadas a un recurso tienden a coincidir entre usuarios, haciendo de esta coincidencia algo especialmente útil en comparación con las etiquetas que encontramos en sistemas de etiquetado simple.

En un sistema de etiquetado colaborativo, como ejemplo, un usuario podría etiquetar este trabajo como *etiquetado-social*, *investigación* y *tesis*, mientras que otro usuario podría utilizar las etiquetas *etiquetado-social*, *marcadores-sociales*, *doctorado* y *tesis* para anotarlo. El comportamiento de los usuarios puede diferir de forma considerable en estos sistemas, donde la

acumulación de sus anotaciones se suele considerar como consenso. Por ejemplo, el resultado de la acumulación mediante suma de las anotaciones de arriba sería el siguiente: tesis (2), etiquetado-social (2), marcadores-sociales (1), doctorado (1) e investigación (1).

En esta tesis, analizamos y estudiamos las anotaciones provistas por usuarios en sistemas de etiquetado social. Presentamos un estudio con el fin de sacar el máximo partido de ellas, con vistas a mejorar el rendimiento de una tarea de clasificación de recursos. Concretamente, nos centramos en el análisis de la utilidad de las folksonomías generadas por usuarios como aproximación a una organización parecida a las taxonomías creadas por expertos. En este contexto, estudiamos diferentes representaciones basadas en el uso de anotaciones sociales, en busca de una aproximación que se parezca a la clasificación provista por expertos en la mayor media posible. Nos centramos en obtener el máximo de las etiquetas sociales, tanto buscando la mejor representación, como midiendo el impacto que puede tener en este sentido la distribución de las etiquetas sobre recursos, marcadores y usuarios. Finalmente, también estudiamos la aplicación de técnicas vanguardistas de análisis del comportamiento de los usuarios en estos sistemas, con el fin de detectar usuarios cuyas anotaciones estén más próximas a la clasificación creada por expertos.

## D.2 Objetivos

El objetivo principal de esta tesis se centra en aportar nuevo conocimiento sobre el uso apropiado de la gran cantidad de datos que se pueden encontrar en los sistemas de etiquetado social. Dado el interés en clasificar recursos, y la falta de datos representativos, nos centramos en analizar en qué medida y de qué manera las etiquetas sociales pueden mejorar la tarea de clasificación de recursos. Al comienzo de este trabajo comprobamos que no había investigaciones que abordaran este problema; por lo tanto, nos motivó a llevar a cabo esta investigación. Con este fin, hemos definido el siguiente planteamiento del problema, el cual resume el objetivo principal de esta tesis:

### Planteamiento del Problema

*¿Cómo se pueden aprovechar las anotaciones provistas por usuarios en sistemas de etiquetado social de forma que se obtenga una clasificación de recursos más precisa?*

## D.3 Metodología

La metodología de investigación seguida a lo largo del trabajo se compone de las siguientes 6 partes:

1. Revisión y lectura del estado del arte, así como estudiar y comprender detalladamente el funcionamiento de los sistemas de etiquetado social.
2. Búsqueda de un clasificador SVM apropiado para llevar a cabo la investigación.
3. Búsqueda de colecciones existentes con información extraída de sistemas de etiquetado social. Como no encontramos ninguno que cumpliera nuestros requisitos, hubo que crear tres colecciones de gran escala en su lugar.
4. Pensar y proponer aproximaciones que se ajusten a la tarea de clasificación basada en etiquetas sociales.
5. Evaluación de las aproximaciones propuestas.
6. Realización de un riguroso análisis de los resultados, con el fin de llegar a unas conclusiones sólidas.
7. Presentación de resultados parciales en congresos y talleres nacionales e internacionales, con el fin de obtener comentarios y sugerencias de otros investigadores.
8. Resumir en esta tesis la investigación, aportaciones, y conclusiones alcanzadas a lo largo de todo el trabajo.

Del paso 4 al 6, se realizó un proceso iterativo, realizándose dichos pasos de forma repetida varias veces.

## D.4 Estructura de la Tesis

Esta tesis está compuesta de 8 capítulos. A continuación resumimos brevemente el contenido de cada uno de ellos:

### Capítulo 1 en la página 21

#### Introducción

Presentamos la motivación para el estudio del uso de anotaciones sociales para clasificación de recursos. Formalizamos el problema y motivamos la necesidad de realizar dicho estudio.

### Capítulo 2 en la página 33

#### Trabajo Relacionado

Ofrecemos un resumen de los trabajos previos en el campo de investigación. Resumimos los avances en campos relacionados, tanto en cuanto al uso de anotaciones sociales, como en cuanto a la clasificación de recursos.

**Capítulo 3 en la página 47****Máquinas de Vectores de Soporte para Clasificación a Gran Escala**

Realizamos un estudio de diferentes aproximaciones SVM para resolver el problema de la clasificación de grandes colecciones de recursos sobre taxonomías multiclase. Damos con la mejor aproximación SVM para estos casos, y la utilizamos a lo largo del trabajo para realizar las tareas de clasificación.

**Capítulo 4 en la página 59****Creación de Colecciones de Etiquetado Social**

Describimos y analizamos en detalle las colecciones de etiquetado social utilizadas en esta tesis. Detallamos el proceso de generación de dichas colecciones, y analizamos las principales características de sus correspondientes folksonomías.

**Capítulo 5 en la página 75****Representando la Acumulación de Etiquetas**

Proponemos y evaluamos diferentes representaciones de recursos que emplean información de etiquetas sociales para la tarea de clasificación de recursos. Estudiamos la utilidad de las etiquetas sociales en comparación a otras fuentes de datos, y proponemos una representación que saca el máximo partido de ellas. También abordamos el problema combinando las etiquetas sociales con las otras fuentes de datos disponibles para obtener un mejor rendimiento.

**Capítulo 6 en la página 95****Analizando la Distribución de Etiquetas para Clasificación de Recursos**

Abordamos la idea de considerar la representatividad de las etiquetas dentro de una colección de anotaciones de un sistema de etiquetado social. Estudiamos la aplicación de funciones de pesado adaptadas a estos sistemas, y analizamos su adecuación teniendo en cuenta las configuraciones de cada sistema.

**Capítulo 7 en la página 111****Analizando el Comportamiento de Usuarios para la Clasificación**

Exploramos el efecto que puede tener el comportamiento de usuarios en sistemas de etiquetado social con vistas a una tarea de clasificación de recursos. Basándonos en trabajos previos que sugieren la existencia de ciertos usuarios que tienden a categorizar recursos, estudiamos si realmente se ajustan en mayor medida a la clasificación de recursos.

**Capítulo 8 en la página 125****Conclusiones y Trabajo Futuro**

Resumimos y comentamos las principales conclusiones y aportaciones del

trabajo. Presentamos las respuestas a las preguntas de investigación formuladas al inicio, y planteamos el trabajo futuro.

Además, la tesis contiene los siguientes apéndices al final, con información adicional y resúmenes en otros idiomas:

**Apéndice A en la página 143**

**Resultados Adicionales**

Presentamos algunos resultados adicionales, los cuales decidimos no incluir en el contenido de la tesis por claridad, pero que merece la pena mostrar ya que ayudan a demostrar y entender algunas conclusiones.

**Apéndice B en la página 145**

**Palabras Clave y Definiciones**

Listamos los términos más relevantes relacionados con los sistemas de etiquetado social y proporcionamos definiciones detalladas.

**Apéndice C en la página 147**

**Lista de Acrónimos**

Listamos los acrónimos utilizados a lo largo de este trabajo e indicamos a qué se refieren.

**Apéndice D en la página 149**

**Resumen**

Resumen del contenido de este trabajo en castellano.

**Apéndice E en la página 167**

**Laburpena (Resumen en euskera)**

Resumen del contenido de este trabajo en euskera.

## **D.5 Preguntas de Investigación Resueltas**

**Pregunta de Investigación 1**

*¿Qué tipo de clasificador SVM debería utilizarse para llevar a cabo este tipo de tareas de clasificación: un clasificador multiclase nativo, o una combinación de clasificadores binarios?*

Se ha demostrado una clara superioridad de los clasificadores SVM multiclase nativos sobre las otras aproximaciones que combinan clasificadores binarios. Los resultados muestran que basarse en un conjunto de clasificadores binarios no es una buena opción cuando se trata de taxonomías multiclase. Por lo tanto, los clasificadores multiclase nativos, que consideran todas las clases al mismo tiempo y tienen más conocimiento de la tarea completa, funcionan mejor para estos casos.

**Pregunta de Investigación 2**

*¿Qué método de aprendizaje rinde mejor para este tipo de tareas de clasificación: uno supervisado o uno semisupervisado?*

Los métodos semisupervisados podrían rendir mejor cuando el subconjunto etiquetado es muy pequeño, pero los métodos supervisados, computacionalmente menos costosos, consiguen un rendimiento muy similar con unas pocas instancias más etiquetadas. Por lo tanto, hemos mostrado también que, a diferencia de las tareas de clasificación binarias como ya demostró [Joachims \(1999\)](#), un método supervisado obtiene unos resultados muy similares a los de un semisupervisado para estos casos de colecciones grandes y multiclase. Parece razonable pensar que predecir la clase de las instancias no etiquetadas es mucho más difícil con el incremento del número de clases y, por tanto, el incremento de errores en las predicciones se refleja también en la fase de aprendizaje del clasificador.

Basándonos en estas conclusiones, decidimos utilizar un clasificador SVM multiclase supervisado a lo largo de esta tesis.

**Pregunta de Investigación 3**

*¿Cómo afecta la configuración de los sistemas de etiquetado social en las anotaciones de los usuarios y las folksonomías resultantes?*

Con este fin, hemos analizado diversas características que se encuentran en la configuración de los sistemas de etiquetado social. Entre las características analizadas, hemos mostrado el gran impacto de las sugerencias en el etiquetado, lo cual altera de forma considerable la folksonomía resultante. En los sistemas de etiquetado social que hemos estudiado, todos presentan alguna característica diferente en este aspecto:

- **Sugerencias basadas en recursos (Delicious):** cuando el sistema sugiere etiquetas asignadas por otros usuarios al recurso que se está guardando, se reduce la probabilidad de utilizar nuevas etiquetas que aporten nueva información. En este caso, los usuarios dedican poco esfuerzo a pensar por ellos mismos, y prefieren basarse en las sugerencias provistas por el sistema.
- **Sugerencias basadas en la personomía (GoodReads):** cuando el sistema sugiere etiquetas que el mismo usuario ha utilizado previamente, el vocabulario de su personomía tiende a ser mucho más reducido. No obstante, los usuarios no saben qué es lo que otros han anotado sobre cada recurso, y por tanto es muy probable que aporten nuevas etiquetas que anteriormente no se habían anotado sobre el recurso.
- **Ausencia de sugerencias (LibraryThing):** cuando el sistema no sugiere etiquetas al usuario, el vocabulario de su personomía tiende a ser mayor, así como las etiquetas asignadas a cada recurso son más diversas.

**Pregunta de Investigación 4**

*¿Cuál es la mejor manera de acumular las anotaciones de los usuarios sobre un recurso con el fin de obtener una representación?*

Hemos demostrado que es mejor tener en cuenta todas las etiquetas anotadas sobre un recurso que basarse sólo en aquéllas que han sido anotadas por más usuarios. Las etiquetas más anotadas han demostrado ser las más importantes, y aportan la información más relevante sobre la temática del recurso. No obstante, las etiquetas menos populares también pueden ser útiles en menor medida, aportando otro tipo de información útil que mejora el rendimiento del clasificador.

En cuanto a los pesos que se asignan a las etiquetas al representar el recurso, los mejores resultados se obtienen considerando el número de usuarios que anotan cada etiqueta. El uso de este valor ha producido los mejores resultados en nuestros experimentos, superando a otras aproximaciones que ignoran estos pesos, y demostrando que no hace falta considerar el número total de usuarios que anota el recurso.

Por lo tanto, a partir de nuestros experimentos, concluimos que la mejor representación es aquélla que aprovecha todas las etiquetas, asignando como peso el número de usuarios que las ha anotado.

**Pregunta de Investigación 5**

*A pesar de la utilidad de las etiquetas sociales para estas tareas, ¿merece la pena considerar otras fuentes de datos como el contenido de los recursos para mejorar aún más los resultados?*

Utilizando técnicas de combinación de clasificadores, los cuales consideran las predicciones de diferentes clasificadores, hemos demostrado que las etiquetas aportan criterios fiables a tener en cuenta. Estos criterios son muy útiles para combinar dichas etiquetas con otras fuentes de datos. No obstante, no todas las fuentes de datos son útiles para combinar, y se deben seleccionar con cautela las que obtienen unos resultados sólidos y, además, ofrecen unas predicciones fiables. Cuando las fuentes de datos se escogen de manera apropiada, la mejora de rendimiento es considerable.

**Pregunta de Investigación 6**

*¿Son las etiquetas sociales también útiles y suficientemente específicas para clasificar recursos en categorías a nivel más bajo?*

Hemos analizado la utilidad de las etiquetas sociales para la clasificación sobre dos niveles diferentes de las taxonomías. Además de las categorías de más alto nivel, también hemos explorado la clasificación sobre categorías del segundo nivel, más precisas. En este aspecto, los resultados usando etiquetas sociales han sido superiores a los obtenidos con otras fuentes de datos para aquellos sistemas



de etiquetado social que animan a los usuarios a aportar anotaciones (Delicious y LibraryThing). La superioridad es muy clara en estos casos, sobre todo para Delicious, donde la diferencia es aún mayor cuando se trata del segundo nivel taxonómico. Esta diferencia es muy similar para LibraryThing. Por último, las etiquetas de GoodReads no superan a las otras fuentes de datos, ni siquiera para el primer nivel, ya que el sistema no anima a los usuarios a anotar los libros, con lo que muchos de los marcadores se quedan sin etiquetas.

Estos descubrimientos arrojan una conclusión diferente a la que dan [Noll and Meinel \(2008a\)](#), donde los autores lanzan la hipótesis de que las etiquetas sociales podrían no ser útiles para niveles más bajos de las taxonomías, y que deberían utilizarse otros tipos de datos para estos casos.

### Pregunta de Investigación 7

*¿Podemos tener en cuenta la distribución de etiquetas a lo largo de la colección para así medir la representatividad general de la etiqueta?*

A través de la experimentación llevada a cabo en esta tesis, hemos demostrado la utilidad de considerar las distribuciones de etiquetas a lo largo de la colección, por medio de una función de pesado inversa como la ofrecida por IDF. Estas funciones han servido para determinar la representatividad de las etiquetas para cada colección, con el fin de mejorar el rendimiento de la tarea de clasificación de recursos. No obstante, hemos mostrado que la configuración del sistema de etiquetado social tiene mucho que ver con esas distribuciones. Entre las características en la configuración de los sistemas, se ha visto que las sugerencias basadas en los recursos influyen en gran medida la estructura de las folksonomías resultantes. Aquellos sistemas que sugieren etiquetas al usuario, basándose en anotaciones previas sobre el recurso, producen unas distribuciones de etiquetas muy diferentes a aquéllos que no sugieren etiquetas y dejan a los usuarios que hagan su propia elección. Esta característica ha sido determinante también para la aplicación con éxito de las funciones de pesado sobre estas distribuciones.

Hemos descubierto que las funciones de pesado de etiquetas propuestas superan claramente a la aproximación basada en TF cuando el sistema no sugiere etiquetas basadas en los recursos (es decir, en LibraryThing y GoodReads), tanto cuando se utilizan por sí solas, como cuando se combina con otras fuentes de datos. En realidad, es mejor considerar simplemente la aproximación basada en etiquetas que combinarla con otras fuentes de datos, ya que por sí sola ofrece los mejores resultados, los cuales no son mejorados cuando se combinan.

No obstante, cuando el sistema sugiere etiquetas basadas en el recurso, las folksonomías generadas son muy diferentes al resto. Esto afecta a las distribuciones de etiquetas en gran medida y, por lo tanto, a las funciones de pesado que hemos estudiado. Debido a ello, el uso de funciones de pesado de etiquetas obtiene peores resultados que no tenerlos en cuenta, y necesitan ser combinadas con

otras fuentes de datos para funcionar mejor. En este último caso, pueden llegar a mejorar a la aproximación basada en TF, gracias a las buenas predicciones que aporta, que ayuda a alimentar de forma adecuada la combinación de clasificadores.

#### **Pregunta de Investigación 8**

*¿Cuál es la mejor aproximación para establecer la representatividad de las etiquetas en la colección?*

Entre las funciones de pesado que hemos estudiado, aquélla que se basa en las frecuencias en marcadores ha demostrado ser la mejor para los sistemas sin sugerencias de etiquetas basadas en recursos. En estos casos, IBF es la mejor opción, seguida por IRF e IUF. Todos ellos superan con claridad a TF, tanto cuando se utilizan por sí solas, como cuando se combinan con otras fuentes de datos.

Por otro lado, cuando el sistema sugiere etiquetas basadas en el recurso es mejor basarse en la frecuencia en usuarios. IUF funciona mejor que IBF e IRF en estos casos, debido a la importancia de aquellos usuarios que tienden a escoger sus propias etiquetas en lugar de basarse en las sugerencias. Aunque ni siquiera IUF supera a TF, cuando se combina con otras fuentes de datos llega a ser la mejor opción. No obstante, los resultados de este último caso son sólo ligeramente superiores a los obtenidos por la combinación que utiliza TF, por lo que cualquiera de ellas podría emplearse para llegar a obtener unos resultados parecidos.

#### **Pregunta de Investigación 9**

*¿Podemos discriminar diferentes perfiles de usuario de manera que encontremos un subconjunto de usuarios que proporciona anotaciones que se ajustan en mayor medida a la tarea de clasificación?*

Hemos demostrado que dicho tipo de usuario, llamado Categorizador, en realidad existe. Según nuestros experimentos, esto es verdad sobre todo cuando se trata de sistemas sin sugerencias de etiquetas como en LibraryThing, donde la clasificación de recursos realizada utilizando etiquetas de los usuarios Categorizadores obtiene mejores resultados. Cuando las sugerencias existen, la detección de usuarios que se adecúan a la tarea se complica, como hemos demostrado que ocurre con GoodReads y Delicious. Sin embargo, la utilización de la medida apropiada puede producir una selección exitosa de usuarios que se ajustan a las características de un Categorizador.

#### **Pregunta de Investigación 10**

*¿Cuáles son las características que identifican a un usuario como apropiado para la tarea de clasificación de recursos?*

De las dos características que hemos considerado en este trabajo, hemos visto que si se diferencian los usuarios por su nivel de verbosidad, se puede encontrar

un conjunto de usuarios que se ajustan más a la tarea de clasificación. Por otra parte, hemos visto que separando usuarios por la diversidad de su vocabulario no se consigue una buena discriminación para este fin, sino para encontrar otro tipo de usuarios llamados Descriptores. Además de esto, hemos visto que aquéllos usuarios que no utilizan datos descriptivos en sus anotaciones ofrecen etiquetas que se ajustan mejor a la clasificación de recursos.

## D.6 Principales Contribuciones

La idea novedosa de este trabajo de investigación se basa en la utilización de anotaciones sociales para enriquecer una tarea de clasificación de recursos. Hasta donde nosotros sabemos, el primer trabajo de investigación que llevó a cabo experimentos con tareas de clasificación reales fue nuestro primer trabajo en este campo (Zubiaga et al., 2009d). Previamente, sólo Noll and Meinel (2008a) habían realizado un análisis estadístico que comparaba etiquetas sociales con una clasificación hecha por expertos. Teniendo en cuenta la carencia de trabajos en el área, la investigación recogida en esta tesis aporta nuevo conocimiento hacia el uso y modo de representación apropiados de etiquetas sociales para la clasificación de recursos. Concretamente, nuestras aportaciones principales al área de investigación son las siguientes:

- Hemos creado 3 colecciones de gran escala que incluyen tanto etiquetas sociales como información de la categoría correspondiente para una serie de recursos. Éstas pueden considerarse como unas de las mayores colecciones utilizadas en el área de investigación y, por lo que nosotros sabemos, las mayores utilizadas para clasificación de recursos. Algunas de estas colecciones, junto con otras más pequeñas que hemos creado a lo largo del trabajo, se han hecho públicas para fines de investigación<sup>11</sup>. Entre otros, Godoy and Amandi (2010) y Strohmaier et al. (2010b) han utilizado alguna de nuestras colecciones para su investigación.
- Nuestro trabajo es el primero que compara diferentes representaciones de recursos usando etiquetas sociales. Además, es el primer trabajo que realiza tareas de clasificación comparando etiquetas sociales con otros tipos de fuentes de datos. Hemos demostrado que las etiquetas sociales son también útiles para categorías más precisas de más bajo nivel. Al contrario de lo que indican que Noll and Meinel (2008a), donde los autores realizan un estudio estadístico con el que concluyen que las etiquetas sociales podrían no ser útiles para categorías más precisas, hemos demostrado que son aún más útiles que para categorías más generales.

---

<sup>11</sup><http://nlp.uned.es/social-tagging/datasets/>

- Hemos analizado las distribuciones de etiquetas sociales en folksonomías, y hemos realizado un riguroso estudio de cómo la configuración de un sistema de etiquetado social afecta tales distribuciones. En este aspecto, hemos adaptado funciones de pesado basadas en la consolidada TF-IDF al ámbito del etiquetado social y las folksonomías.
- Hemos mostrado la existencia de un grupo de usuarios, llamados Categorizadores, cuyas anotaciones se parecen más que las de otro grupo de usuarios, llamados Descriptores, a la clasificación hecha por expertos. Aunque la aproximación para diferenciar Categorizadores y Descriptores ya estaba consolidada de previos trabajos, en éste hemos llevado a cabo la tarea de demostrar que los Categorizadores se ajustan más a la clasificación de recursos.

La utilización de anotaciones sociales para el beneficio de tareas de clasificación de recursos era una línea de investigación nueva al comienzo de esta tesis. Sin embargo, el crecimiento en el interés de los investigadores sobre contenidos generados por usuarios en medios de comunicación social, y concretamente en los sistemas de etiquetado social, ha ocasionado recientemente la aparición de numerosos trabajos en el área. Junto con este crecimiento, más investigadores han mostrado su interés en utilizar anotaciones sociales para clasificación de recursos, y el número de trabajos relacionados ha aumentado considerablemente. [Godoy and Amandi \(2010\)](#), por ejemplo, presentan un estudio de clasificación basada en etiquetas que se inspira en un trabajo nuestro ([Zubiaga et al., 2009d](#)).

## D.7 Trabajo Futuro

La utilización de anotaciones sociales para la clasificación de recursos es un campo de investigación que está aún en sus inicios, y se ha realizado relativamente poco trabajo hasta el momento. El trabajo presentado en esta tesis concluye con la manera de representar etiquetas sociales en busca de una clasificación de recursos lo más precisa posible. Además, da lugar al planteamiento de diversos trabajos futuros.

A lo largo de esta tesis hemos considerado cada etiqueta como un símbolo diferente, sin tener en cuenta su significado semántico. En este aspecto, nuestros planes para trabajo futuro incluyen el análisis del significado de las etiquetas para tratar de descubrir palabras sinónimas y relaciones entre ellas. Bien utilizando técnicas de procesamiento de lenguaje natural, o bien mediante aproximaciones semánticas, esto podría ayudar a entender el significado de cada etiqueta, pudiendo explorar más allá el conocimiento que aportan las folksonomías.

Las tres funciones de pesado que hemos empleado en el Capítulo 6 se basan en la conocida TF-IDF, que fue diseñada inicialmente para colecciones de texto.

Pensamos que probar otras funciones de pesado, así como explorar la posible definición de una nueva función que se ajuste a las necesidades de estas estructuras sociales, pueden resultar en interesantes aportaciones como trabajo futuro. Esto ayudaría sobre todo para sistemas que dan sugerencias de etiquetas basadas en recursos, como pasa con Delicious, donde las funciones de pesado que hemos experimentado no han dado buenos resultados.

---



## Laburpena (Basque Summary)

*“Hizkuntza bat ez da galtzen ez dakitenek ikasten ez dutelako, dakitenek erabiltzen ez dutelako baizik.”*

— Joxean Artze

# Baliabideen Sailkapenerako Folksonomien Ustiapena

Tesi honetan baliabideen sailkapenaren gainean dihardugu, eguneroko bizitzan hain garrantzitsua eta ohikoa den ataza bat landuz, liburuak katalogatzea edo bideoak antolatzea izan daitekeen bezalaxe. Ataza burutzeko, etiketa sozialen sistemetan erabiltzaileek egindako anotazioez baliatzen gara. Webgune hauetan baliabide ezberdinen gainean metadatu ugari eskaini ohi dituzte erabiltzaileek. Orain arte, gutxi dira metadatu hauek helburu honetarako erabili dituzten ikerketa lanak, eta gutxi horiek analisi estatistikoak egitera mugatu dira. Tesi honetan, sistema hauen, bertako erabiltzaileen eta haien anotazioen ezaugarriak aztertzen ditugu, datu-sorta handi hauetaz ahal bezainbeste profitu nahian, eta ahalik eta baliabideen sailkapen automatiko zehatzena lortu asmoz.

## E.1 Motibazioa

Edozein motatako baliabideak aurrez definitutako kategoriatan sailkatzea ohiko ataza da gure eguneroko bizitzan. Baliabideei kategoriak esleitzeak ondoren berreskuratu ahal izateko erraztasunak eskaintzen ditu, bilaketa nahi den kategoriora mugatuz. Esate baterako, liburuzainek gaika antolatu ohi dituzte liburuak ka-

talogoetan. Horrez gain, filmen datubaseak, musika katalogoak eta fitxategi sistematik, besteak beste, kategoriatan antolatu ohi dira baliabideok aurkitzea erraztuz. Era berean, *Yahoo! Directory* eta *Open Directory Project* bezalako web direktorioek kategoriatan antolatzen dituzte web orrialdeak. Web orrialdeak sailkatuta izateak interneteko bilatzaileen funtzionamendua hobe dezake emaitzak erabiltzailearen intereseko kategoriara mugatuz (Qi and Davison, 2009).

Kategorizatze lan hori eskuz egitea, ordea, oso garestia izaten da baliabide sorta handia denean. Adibide gisa, Estatu Batuetako *Library of Congress* liburutegi publikoak 2002an profesionalen katalogatutako liburu bakoitzak 94,58 dolarreko kostua izan zuela adierazi zuen<sup>1</sup>. Urte hartan katalogatu zituzten 291.749 erregistroengatik 27,5 milioitik gora ordaindu behar izan zituzten beraz. Ataza hau zein garestia den ikusita, sailkatzaile automatikoetara pasatzeak alternatiba egokia dirudi eskulana gutxitzeko, betiere katalogoak eguneratuta mantenduz.

Orain arte, sailkatzaile automatiko gehienak baliabideen edukian oinarritu dira informazio iturri gisa, web orrialdeen sailkapenari dagokionean batik bat (Qi and Davison, 2009). Baliabideen edukiek ez dute beti informazio esanguratsua izaten, ordea, eta horrek zaildu egiten du ataza. Gainera, batzutan ez da erraza izaten liburuak eta filmeak bezalako baliabideentzako datu nahikoa lortzea. Horrelako kasuetan zailagoa izaten da edukia errepresentatzea, eta litekeena da edukia erraz prozesatu daitekeen formatu batean ez izatea.

Arazo hauentzako soluzio posible bezala, etiketa sozialen sistemek baliabideei dagozkien metadatuak eskuratzeko modu errazago eta merkeagoa eskaintzen dute. Delicious<sup>2</sup>, LibraryThing<sup>3</sup> eta GoodReads<sup>4</sup> bezalakoek baliabideen inguruan erabiltzaileek definitutako etiketak batzen dituzte. Erabiltzaileek sortutako etiketa hauek baliabideen edukiak deskribatzen dituzten datu esanguratsuak direla frogatu da (Heymann et al., 2008).

Etiketa hauen bitartez, baliabideen sailkapen propio baten antzekoa eskaintzen dute erabiltzaileek. Eta etiketa hauek modu sozialean elkarbanatzen dira komunitatearekin. Sistema hauetan erabiltzaile kopuru handiek parte hartzen dutenez, beraien anotazioak baliabideen gainean batu egiten dira. Ondorioz, erabiltzaile ezberdinen anotazioak baliabideetan batzeko gaitasun horrek are erabilgarriago eta baliagarriago egiten du anotazio horietako bakoitza. Komunitate aktiboetako erabiltzaileek laster-marka, etiketa eta anotatutako baliabide sorta handiak sor ditzakete.

*“Sailkapen individual bakoitzak profesional batek egindakoak baino gutxiago balio du. Baina ugari, mordoxka bat daude”, Joshua Schachter, Delicious-*

<sup>1</sup><http://www.loc.gov/loc/lcib/0302/collections.html>

<sup>2</sup><http://delicious.com>

<sup>3</sup><http://www.librarything.com>

<sup>4</sup><http://www.goodreads.com>



en sortzailea, 2006ko FOWA bilkuran, Londresen (Ingalaterra)<sup>5</sup>.

Etiketa sozialen sistemak baliabideak gorde, antolatu eta bilatzeko tresnak dira, erabiltzaileek hautatutako etiketak baliatuz anotatzea ahalbidetzen dutenak. Gure ustez, anotazio hauek nabarmen hobe dezakete baliabideen sailkapen automatikoa. Erabiltzaileek sortutako anotazio hauek erabilgarri izan litezke baliabideen kategoriaren inguruko informazioa ematen duen informazio iturri gisa.

Baliabide bakoitzaren gainean erabiltzaile askok esleitzen dituzten anotazioak, gure helburu nagusia berauen ekarpenak batzeko modu egokia aurkitzean datza, betiere profesionalek egindako kategorizazioarekiko antzekoa den antolaketa lortuz asmoz. Erabiltzaile askok metadatu kopurua handia esleitzen duenez, gure erronka ahalik eta emaitza onena lortzean datza.

*“Garaiotan datuak eskuratzeko erraza da, baina hauek zentzuz erabiltzea ez da hain erraza”, Danah Boyd, Microsoft Research New England-eko Social Media gaineko ikertzailea, WWW2010 kongresuan, Raleigh, Ipar Karolina (Ameriketako Estatu Batuak)<sup>6</sup>.*

### E.1.1 Baliabideen Sailkapena

Baliabideen sailkapena aurrez definitutako kategoria sorta batean baliabideak antolatzean datzan ataza da. Tesi honetan, Euskarri Bektoredun Makinak darabiltzigu (Support Vector Machines, SVM, [Joachims \(1998\)](#)), sailkapen metodo abangoardista. Sailkapen ataza mota hauek aurrez sailkatutako baliabide sorta batean oinarritzen dira, berau sailkatzaileak behar duen ezagutza eraikitzeke baliatzen delarik.

Baliabideen sailkapen ataza batek ezaugarri ezberdinak izan ditzake. Alde batetik, sistemaren ikasketa metodoari dagokionean, *gainbegiratu*a dela esaten da ikasteko erabilitako baliabide guztiak aurrez sailkatuta daudenean, eta *erdi-gainbegiratu*a dela, ostera, sailkatu gabeko baliabideen gainean egindako aurreikuspenak ere ikasteko erabiltzen direnean. Bestalde, kategoria kopuruari dagokionean, sailkapena *bitarra* izan daiteke, bi kategoria baino ezin direnean esleitu, edo *kategoria-anitza*, hiru edo kategoria gehiago daudenean. Lehena iragazte sistemetarako erabili ohi da, bigarrena taxonomia handiegoekin erabiltzen delarik, adibidez, gaikako sailkapena.

Baliabideen kolezio handien gaikako sailkapena burutzeko, Web-eko orrialdeak edo liburutegietako liburuak izan daitezkeen bezalaxe, taxonomiak bi kategoria baino gehiagokoak izan ohi dira, eta aurrez sailkatutako baliabide kopurua oso murrizta izaten da. Beraz, interesgarria deritzogu bai teknika erdi-gainbegiratuak eta bai kategoria-anitzak kontuan hartu eta aztertzea, ataza hauek

<sup>5</sup><http://simonwillison.net/2006/Feb/8/summit/>

<sup>6</sup><http://www.danah.org/papers/talks/2010/WWW2010.html>

burutzeko aukerarik onena zein den jakin ahal izateko.

Tesi honetan, SVM algoritmoan oinarritzen diren hainbat metodoren analisia proposatzen dugu, ataza hauekiko duten aproposasuna aztertuz. Metodo hauen artean teknika kategoria-anitz ezberdinak aztertzen ditugu, eta baita teknika gainbegiratu zein erdi-gainbegiratuak ere.

### E.1.2 Anotazio Sozialak

Etiketa sozialen sistemek baliabide gogokoenak (web orrialdeak, filmeak, liburuak, argazkiak edo musika, besteak beste) gorde eta anotatzeko aukera eskaintzen diete erabiltzaileei, komunitatearekin elkarbanatuz. Anotazio hauek etiketa moduan eman ohi dituzte erabiltzaileek. Etiketatzea baliabideei hitz gakoak edo etiketak esleitzeari deritzo, deskribatzeko zein antolatzeako aukera emanez. Honek ondoren baliabideok bilatzea errazten du, etiketa horiek bilaketa gako bezala baliatuz. Sistema gehienetan ez daude aurrez definitutako etiketak, eta beraz nahi duten hitzak hauta ditzakete erabiltzaileek etiketa gisa.

*“Etiketatzek interfazearekin zerikusi handia du - jendeak gauzak gogoratzeko modu bat, gorde zuten unean zertan pentsatzen ari ziren erakusten duena. Nahiko erabilgarria gogoratzeko, ona deskubritzeko, ikaragarria hedatzeko (non argitaratzen dituztenek ahal bezainbeste etiketa definitzen dituzten kutxa gehiagotan sailkatzeko).”, Joshua Schachter, Delicious-en sortzailea, 2006ko FOWA bilkuran, Londresen (Ingalaterra)*<sup>7</sup>.

Etiketatzeko prozesu honen bitartez folksonomia deritzon egitura sortzen da, erabiltzaileek sortutako baliabideen antolaketa, alegia. Folksonomia *folk* (jendea), *taxi* (sailkapena) eta *nomos* (kudeaketa) hitzen laburtzapena da. Komunitatean oinarritzen den taxonomia bezala ere ezagutzen da folksonomia, non sailkapen mota ez-hierarkikoa den, adituek egindako sailkapen taxonomikoetan ez bezala. Beraz, folksonomiek badute nolabaiteko zerikusia adituek egindako sailkapenekin, baliabideak taldeka sailkatzen baitira era berean.

Anotazio hauek sozialak direla esan ohi da ingurune sozial batean komunitatearekin elkarbanatuz beste guztientzako erabilgarri agertzen direnean. Honek dakarren abantaila nagusia besteek ipinitako etiketak baliatuz bilaketak egin ahal izatea da. Era berean, hauxe da erabiltzaile asko parte hartzera animatzen duen ezaugarrietako bat.

Anotazio guztiak ez dira modu berean elkarbanatzen, ordea. Etiketa sozialen guneak berak baldintza batzuk defini ditzake, baliabide bakoitza nork anota dezakeen mugatuz, batez ere. Honi dagokionean, bi sistema mota ezberdin ditzakegu (Smith, 2008):

<sup>7</sup><http://simonwillison.net/2006/Feb/8/summit/>

- **Etiketen sistema sinpleak:** erabiltzaileek norbere baliabideak etiketa ditzakete (adibidez, argazkiak Flickr-en<sup>8</sup>, bideoak Youtube-n<sup>9</sup> edo albisteak Digg-en<sup>10</sup>), baina inork ezin ditu besteen baliabideak etiketatu. Normalean, baliabidearen egilea bera izaten da etiketatzen duena. Ondorioz, baliabide bakoitza erabiltzaile batek baino ez du etiketatzen. Etiketen sistema sinpleetan erabiltzaile sorta bat ( $U$ ) izaten da, baliabide batzuen ( $R$ ) gainean etiketa sorta bat ( $T$ ) esleitzen duena.  $u_i \in U$  erabiltzaile batek  $r_j \in R$  bere baliabidea  $p$  etiketa kopuru aldagarriaren  $T_j = \{t_{j1}, \dots, t_{jp}\}$  etiketa-sortarekin anotatzen du.  $r_j$  baliabideari esleitutako etiketa-sortak  $T_j$  izaten jarraituko du aurrerantzean, beste inork ezingo baitu anotatu.
- **Etiketen sistema kolaboratiboak:** erabiltzaile askok anotatzen dute baliabide bera, bakoitzak etiketa ezberdin batzuk baliatuz. Erabiltzaile bakoitzak erabilitako etiketa sortak folksonomia txikiago bat sortzen du, pertsonomia deritzona. Sistema hauetan hainbat erabiltzailek etiketatu ohi du baliabide bera. Esate baterako, CiteULike.org, LibraryThing.com eta Delicious etiketatze kolaboratiboan oinarritzen dira, non baliabide bakoitza (artikuluak, liburuak eta URLak, hurrenez hurren) interesgarri deritzon erabiltzaile orok etiketa dezakeen. Etiketatze sistema kolaboratiboak sinpleak baino konplexuagoak dira. Sistema hauetan, erabiltzaile sorta bat ( $U$ ) izaten da, baliabide batzuen ( $R$ ) gainean laster-marka batzuk ( $B$ ) gordetzen dabilena, etiketa sorta batekin ( $T$ ) anotatuz.  $u_i \in U$  erabiltzaile batek  $r_j \in R$  baliabidearen  $b_{ij} \in B$  laster-marka gorde dezake  $p$  etiketa kopuru aldagarriaren  $T_{ij} = \{t_{ij1}, \dots, t_{ijp}\}$  etiketa-sorta baliatuz.  $k$  erabiltzailek  $r_j$  baliabidea gorde eta gero,  $T_j = \{w_{j1}t_{j1}, \dots, w_{jn}t_{jn}\}$  pisudun etiketa-sorta bezala defini daitezke bere anotazioak, non  $w_{j1}, \dots, w_{jn} \leq k$  aldagaiek etiketa bakoitzaren esleipen kopurua adierazten duten. Ondorioz, laster-marka bakoitzak erabiltzaile, baliabide eta etiketa-sorta bana ditu bere baitan:  $b_{ij} : u_i \times r_j \times T_{ij}$ . Erabiltzaile bakoitzak baliabide ezberdinen laster-markak egiten ditu, eta aldi berean baliabide batek erabiltzaile ezberdinek egindako laster-markak izan ditzake. Erabiltzaile baten laster-marketako etiketak bateratzearen emaitza pertsonomia izenez ezagutzen da:  $T_i = \{w_{i1}t_{i1}, \dots, w_{im}t_{im}\}$ , non  $m$  erabiltzaileak dituen etiketa ezberdinen kopurua den.

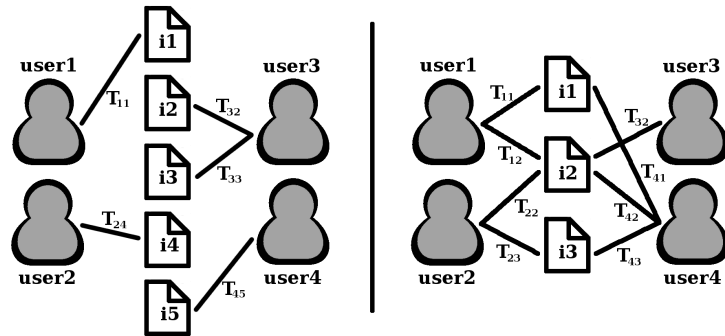
Tesi honetan etiketatze kolaboratiboko sistemekin dihardugu. E.1. Irudiak bi sistema hauen arteko ezberdintasunak erakusten ditu adibide baten bitartez.

Baliabide bera anotatzen duten erabiltzaileen artean etiketak kointziditzeko probabilitatea altua izaten da. Ezaugarri honek bereziki interesgarri egiten du

<sup>8</sup><http://www.flickr.com>

<sup>9</sup><http://www.youtube.com>

<sup>10</sup><http://digg.com>



**Figura E.1:** Etiketatzeko sistema sinplea eta kolaboratiboko sistemaren anotazioen arteko konparazioa.

etiketatzeko kolaboratiboetako erabiltzaileen bateratze hau etiketatzeko sinpleekin alderatuz.

Etiketatzeko sistema kolaboratibo batean, adibidez, erabiltzaile batek lan haxe bera anotatzeko etiketa-sozialak, ikerketa eta tesia etiketak erabil litza-ke, eta beste erabiltzaile batek etiketa-sozialak, markatzaile-sozialak, doktoretza eta tesia etiketak. Erabiltzaile bakoitzaren jarrera oso ezberdina izan daiteke sistema hauetan, eta horrexegatik izaten dira kontuan beraien guztien anotazioak bateratzeko orduan. Adibide bezala aipatutako bi horien anotazioak batuz hurrengoak lortuko genuke: tesia (2), etiketa-sozialak (2), markatzaile-sozialak (1), doktoretza (1) eta ikerketa (1).

Tesi honetan, erabiltzaileek etiketen sistema sozialean egindako anotazioak aztertu eta ikertzen ditugu. Baliabiden sailkapen egoki bat lortzeko etiketa horiengandik etekina ateratzeko ikerketa lana aurkezten dugu. Konkreteki, erabiltzaileek sortutako folksonomiek adituek egindako sailkapen baten antzeko zerbait lortzeko duten erabilgarritasuna aztertzen dugu. Etiketa sozialen errepresentazio ezberdinak aztertzen ditugu bertan, betiere adituek egindako sailkapen horietara hurbildu asmoz. Bereziki, etiketa sozialei etekina ateratzea da gure helburua, bai errepresentazio egokiaren bilatuz, eta baita etiketek baliabide, lastermarka eta erabiltzaile ezberdinetan aurkezten dituzten distribuzioen eraginari erreparatuz. Azkenik, sailkapen ataza batetik gertuago dauden erabiltzaileak bilatzeko ikerketa aurkezten dugu, horretarako erabiltzaileen jarrera antzemateko teknika abangoardistez baliatuz.

## E.2 Helburuak

Tesi honen helburu nagusia etiketa sozialen sistemetan aurkitzen diren anotazio horiei guztiei etekina ateratzeko egin beharreko erabilpen egokiaren inguruan

ezagutza berria zabaltzea da. Baliabideak sailkatzeak duen interesa jakinda, eta berauek errepresentatzeko datu esanguratsuen gabezia kontuan izanda, gure helburua baliabideen sailkapenerako etiketa sozialek lagun dezaketena aztertu, eta erabilpen aproposa egiteko modurik egokiena zein den jakitea da. Lan honen hasieran, hau ikertzen zuen lanik ez zegoela ikusi genuen. Horrek motibatu gintuen ikerketa lan hau aurrera eramatera. Helburu honekin, hurrengo planteamendua egin genuen, lanaren xede nagusia laburbilduz:

#### **Lanaren Planteamendua**

*Nola egin daiteke erabiltzaileek etiketen sistema sozialetan egindako anotazioak usiatzeko baliabideen sailkapen ahalik eta zehatzena lortuz?*

### **E.3 Metodologia**

Tesi hau aurrera eramateko jarraitu den ikerketa metodologiak hurrengo 6 pausoak jarraitu ditu:

1. Artearen egoeraren azterketa eta irakurketa sakona, eta baita etiketa sozialen sistemak ikertu eta funtzionamendua ondo ulertzea ere.
  2. Lana burutu ahal izateko SVM sailkatze egokia aurkitzea.
  3. Etiketen sistema sozialetan oinarrituz sortutako kolekzioak bilatzea. Gure beharrak betetzen zituenik aurkitu ez genuenez, gureak sortzea erabaki genuen, tamaina handiko hiru sortuz.
  4. Etiketa sozialetan oinarrituz, eta sailkapen atazan baliatzeko direla kontuan hartuz, lana aurrera eramateko aproposak diren hurbilketak eta errepresentazioak pentsatzea eta proposatzea.
  5. Proposatutako hurbilketa eta errepresentazioak ebaluatzea.
  6. Emaizen azterketa sakona burutzea, ondorio sendoetara iritsi asmoz.
  7. Egindako lanaren emaitza partzialak kongresu eta tailer nazional eta internazionalen aurkeztu, beste ikertzaileen iritzi eta gomendioak jaso ahal izateko.
  8. Egindako ikerketa lana, ekarpen nagusiak, eta lortutako ondorioak tesi honetan batu eta laburbildu.
4. pausotik 6.era, behin eta berriz errepikatu zen prozesua, pauso horiek behin baino gehiagotan burutuz.

## E.4 Tesiaren Egitura

Tesi honek 8 kapitulu ditu. Jarraian azaltzen da labur-labur kapitulu hauetako bakoitzaren edukia zein den.

### 1. kapitulua 21. orrialdean

#### **Sarrera**

Baliabideen sailkapenerako etiketa sozialak ikertu nahi izateko motibazioa aurkezten dugu. Ataza formalki azaldu, eta ikerketa burutzeko beharra motibatzen dugu.

### 2. kapitulua 33. orrialdean

#### **Erlazionatutako Lana**

Arlo honetan eta erlazionatutakoetan lehenago egindako lanak laburbiltzen ditugu, bai etiketa sozialen erabilpenean, eta baita baliabideen sailkapenean ere.

### 3. kapitulua 47. orrialdean

#### **Euskarri Bektoredun Makinak Neurri Handiko Sailkapenerako**

Taxonomia kategoria-anitzetan baliabide kolekzio handien sailkapenerako SVM hurbilketa ezberdinen analisia aurkezten dugu. Tesian zehar erabiltzeko aproposena den SVM hurbilketa zein den jakitea ahalbidetzen digu ikerketa honek.

### 4. kapitulua 59. orrialdean

#### **Etiketa Sozialen Datu-Sorten Sorkuntza**

Lan honetan guztian zehar erabiltzeko sortu genituen etiketa sozialen kolekzioak zehatz-mehatz deskribatu eta aztertzen ditugu. Kolekzioak sortzeko jarraitutako prozesua azaldu, eta folksonomien ezaugarri nagusiak aztertzen ditugu.

### 5. kapitulua 75. orrialdean

#### **Etiketen Gehikuntzaren Errepresentazioa**

Baliabideen sailkapenerako etiketa sozialen errepresentazio ezberdinak proposatu eta ebaluatzen ditugu. Etiketa sozialek beste datu iturri batzuekin alderatuta baliabideen sailkapenerako duten errendimendua ikertzen dugu, eta ataza burutzeko errepresentazio egokiena zein izan daitekeen aztertzen dugu. Horrez gain, etiketa sozialak beste datu iturriekin nahasten ditugu errendimendua hobetu ahal izateko.

### 6. kapitulua 95. orrialdean

#### **Baliabideen Sailkapenerako Etiketen Distribuzioaren Azterketa**

Etiketa sozialen sistemetako etiketa bakoitzak baliabideen sailkapenerako duen adierazgarritasuna aztertzen dugu. Horretarako, sistema hauetarako

egokitutako pisu-funtzioak erabiltzen ditugu. Gainera, funtzio hauek zenbaterainoko aproposak diren aztertzen dugu, betiere sistema bakoitzaren ezarpenei erreparatuz.

#### 7. kapitulua 111. orrialdean

##### **Sailkapenerako Erabiltzaileen Jarreraren Analisia**

Etiketa sozialen sistemetako erabiltzaileen jarrerak baliabideen sailkapenean izan dezakeen eragina aztertzen dugu. Sailkatzea helburu duten erabiltzaileak existitzen direla dioten aurreko lanetan oinarrituz, erabiltzaile horiek baliabideen sailkapenerako egokiagoak diren aztertzen dugu.

#### 8. kapitulua 125. orrialdean

##### **Ondorioak eta Etorkizunerako Ildoak**

Lanaren ondorio eta ekarpen nagusiak laburbiltzen ditugu. Horrez gain, lanaren hasieran formulatutako galderei erantzun, eta etorkizunerako ildoak aurkezten ditugu.

Horrez gain, tesi honek jarraian azaltzen diren eranskin hauek ere baditu, informazio gehigarria eta beste hizkuntza batzuetako laburpenekin:

#### A. eranskina 143. orrialdean

##### **Emaitza Gehigarriak**

Gehigarri gisa, emaitza lagungarri batzuk aurkezten ditugu, tesiaren parte moduan sartu ez baditugu ere, ondorio batzuk frogatu eta ulertzeko balio dutenak.

#### B. eranskina 145. orrialdean

##### **Hitz Nagusiak eta Definizioak**

Etiketa sozialen sistemekin zerikusia duten hainbat hitzen definizioa ematen dugu.

#### C. eranskina 147. orrialdean

##### **Akronimoen Zerrenda**

Lanean zehar erabilitako akronimoen eta berauen esanahien zerrenda aurkezten da.

#### D. eranskina 149. orrialdean

##### **Resumen (Gaztelerazko Laburpena)**

Lan honen edukiaren gaztelerazko laburpena.

#### E. eranskina 167. orrialdean

##### **Laburpena**

Lan honen edukiaren euskarazko laburpena.

## E.5 Ebatzitako Ikerketa Galderak

### 1. Ikerketa Galdera

*Zein SVM sailkatzaile mota erabili beharko litzateke sailkapen ataza hauek burutzeko: jatorrizko klase-anitzeko sailkatzailea, ala sailkatzaile bitarren konbinazio bat?*

Jatorrizko kategoria-anitzeko SVM sailkatzaile bat erabiltzea sailkatzaile bitarra bateratzea baino askoz aproposagoa dela erakutsi dugu. Gure emaitzek argi eta garbi erakutsi dute kategoria-anitzeko taxonomien kasuan ez dela aukera aproposa sailkatzaile bitarretan oinarritzea. Ondorioz, jatorrizko kategoria-anitzeko sailkatzaileak, kategoria guztiak aldi berean kontuan hartuz ataza osoa hobeto ezagutzen dutenak, egokiagoak dira errendimendu hobearen lortzeko.

### 2. Ikerketa Galdera

*Zein ikasketa motak ematen du errendimendu hobearen sailkapen ataza hauek burutzeko: gainbegiratu batek, ala erdi-gainbegiratu batek?*

Teknika erdi-gainbegiratuak emaitza hobeak eskura ditzakete aurrez sailkatutako baliabide sorta oso-oso txikia denean, baina teknika gainbegiratuak antzeko errendimendua lortzen dute baliabide gehixeago kontuan hartuz. Horrez gain, teknika gainbegiratuak konputazio aldetik gutxiago exijitzen dute. Beraz, [Joachims \(1999\)](#) egileak sailkapen bitarrerako erakutsitakoaren aurkakoa erakutsi dugu, ataza kategoria-anitzetarako teknika gainbegiratu eta erdi-gainbegiratuak oso antzerakoak direla, alegia. Zentzuzkoa dirudi kategoria kopurua handitu ahala zailagoa izatea teknika erdi-gainbegiratuaren ikasketa behar bezala burutzea, gaizki sailkatutako baliabideek zarata gehitzen baitute ikasketa prozesuan.

Ondorio hauei eutsiz, tesian zehar kategoria-anitzeko SVM gainbegiratua erabiltzea erabaki genuen.

### 3. Ikerketa Galdera

*Nola eragiten dute etiketa sozialen sistemetako ezarpenek bertako erabiltzaileen anotazioetan eta ondorioz sortutako folksonomietan?*

Hau jakiteko, etiketa sozialen sistemetako ezarpenen ezaugarri ezberdinak aztertu ditugu. Aztertutako ezaugarrien artean, etiketak gomendatzeak duen garrantzia nabaritu dugu, folksonomien egitura nabarmen eragiten baitu. Aztertutako etiketa sozialen sistemek ezarpen ezberdinak dituzte gomendioei dagokienean:

- **Baliabidean oinarritutako gomendioak (Delicious):** baliabide bat etiketatzerako orduan, sistemak baliabide horretan beste erabiltzaile batzuek definitutako etiketak gomendatzen dituenean, erabiltzaileak etiketa berriak



definitzeko probabilitatea izugarri jaisten da, gehienetan gomendioetan oinarritzen baitira. Kasu honetan, erabiltzaileek esfortzu txikia egiten dute etiketa berriak pentsatzen, eta gomendioei kasu egitea nahiago izaten dute.

- **Pertsonomian oinarritutako gomendioak (GoodReads):** baliabide bat etiketatzerako orduan, sistemak erabiltzaile horrek aurrez beste baliabide batzuetan ipinitako etiketak gomendatzen dituenean, erabiltzailearen etiketa kopurua izugarri murrizten da. Erabiltzailearen berbategia askoz txikiagoa izaten da beraz. Hala eta guztiz ere, erabiltzaileek ez dakite beste batzuek zein etiketa ipini dizkioten baliabideari, eta baliabidearekiko berriak diren etiketak definitzeko probabilitatea mantendu egiten da.
- **Gomendiorik gabe (LibraryThing):** baliabide bat etiketatzerako orduan, sistemak etiketarik gomendatzen ez duenean, erabiltzailearen berbategia hazi egiten da, eta baliabide bakoitzean etiketa berriak definitzeko probabilitatea mantendu egiten da.

#### 4. Ikerketa Galdera

*Zein da baliabide baten gainean erabiltzaileek egindako anotazio guztiak adierazpen bakarrean bateratzeko modurik egokiena?*

Erabiltzaile gehienek anotatu dituzten etiketa gutxi batzuetan oinarritu baino, etiketa guzti-guztiak kontuan hartzea merezi duela erakutsi dugu. Gehien anotatutakoak dira garrantzitsuenak, eta baliabidea zeren ingurukoa den gehien adierazten dutenak dira. Hala eta guztiz ere, erabiltzaile gutxik anotatutakoek ere badute nolabaiteko adierazgarritasuna, beste neurri batean bada ere, eta sailkatzailearentzako baliagarria den informazioa eskaintzen dute.

Baliabideen errepresentazioa egiterakoan etiketei emandako pisuei dagokionean, etiketa bakoitza definitu duen erabiltzaile kopurua pisu bezala erabiltzearena da emaitza onenak ematen dituen. Erabiltzaile kopurua alde batera utzi, edo baliabidea anotatu duen erabiltzaile guztien kopurua kontuan izatea bezalako beste hurbilketa batzuk gaingitu ditu aurrekoak.

Laburbilduz, aurkitu dugun errepresentazio egokiena etiketa guztiak erabili, eta etiketa bakoitza erabiltzaile kopuruaren arabera pisatzearena da.

#### 5. Ikerketa Galdera

*Etiketa sozialek ataza hauetarako duten balioaz gainera, merezi al du baliabidearen barne edukia bezalako beste datu iturri batzuk kontuan hartzea emaitzak are gehiago hobetzeko?*

Sailkatzaile ezberdinen aurreikuspenak elkartzen dituzten sailkatzaile bateratuetan oinarrituz, etiketek kontuan hartu beharreko iritziak ematen dituztela

erakutsi dugu. Iritzi hauek oso baliagarriak dira etiketa sozialak beste datu iturriekin bateratzeko. Edonola ere, datu iturri guztiak ez dira lagungarriak bateratzerako orduan. Aukeratutako datu iturriak nahikoa sendoak izan behar dira, aurreikuspen iritzi aproposak eman ditzaten. Datu iturriak ondo aukeratzen direnean, baina, errendimendua nabarmen hobe daiteke.

## 6. Ikerketa Galdera

*Baliabideak maila baxuagoko kategoria zehatzagoetan sailkatu ahal izateko nahikoa erabilgarriak eta zehatzak dira etiketa sozialak?*

Baliabideen sailkapenerako etiketa sozialean erabilgarritasuna taxonomien bi mailatan aztertu dugu. Goi-mailako kategoriez gainera, bigarren mailako kategoria zehatzagoekin ere egin dugu ikerketa. Guneak erabiltzaileak anotatzera animatzen dituenean (Delicious eta LibraryThing-en), etiketa sozialek beste datu iturriek baino emaitza hobeak lortzen dituzte. Etiketek askogatik gainditzen dituzte beste datu iturriak kasu hauetan, Delicious-en batez ere, bigarren mailako emaitzek abantaila askoz garbiagoa erakusten baitute. Ezberdintasun hau LibraryThing-en ere gertatzen da. Azkenik, GoodReads-eko etiketek ez dituzte beste datu iturriak gainditzen, ezta goi-mailako kategorietan ere, sistemak ez di-tuelako erabiltzaileak etiketatzen animatzen, eta horrela anotazio gutxiago egiten direlako.

Gure emaitza hauek [Noll and Meinel \(2008a\)](#) egileen hipotesia deusezten dute. Beraiek egindako analisi estatistikoan, etiketek kategoria zehatzagoetan sailkapenak egiteko balioko ez zutela uste zuten, eta horretarako beste datu iturri batzuk erabili beharko liratekeela.

## 7. Ikerketa Galdera

*Etiketen adierazgarritasuna neurtzera bidean, kolekzioan zehar etiketek duten distribuzioa kontuan har al daiteke?*

Etiketen distribuzioak kontuan hartzea, IDFn oinarritutako pisu-funtzio batean oinarrituz, baliabideen sailkapenerako etiketen adierazgarritasuna zehazteko interesgarria dela erakutsi dugu. Etiketa sozialen sistemaren ezarpenek, ordea, zerikusi handia dute distribuzio hauekin. Guneen ezarpenen artean baliabideetan oinarritutako gomendioek garrantzia handia dutela ikusi dugu, folksonomiaren egitura erabat aldatzen baitute. Gomendio hauek dituzten sistemek oso distribuzio ezberdinak aurkezten dituzte. Honen arabera, pisu-funtzioen erabilgarritasuna jakin daiteke.

Gure sailkapen esperimentuetan ikusi ahal izan dugu pisu-funtzioek TF gainditzten dutela baliabideetan oinarritutako gomendioak existitzen ez direnean, hau da, LibraryThing eta GoodReads-en, bai bakarrik erabilita, eta baita beste datu iturri batzuekin elkartzekoan ere. Halaber, aproposagoa da berauek bakarrik

erabiltzea, beste datu iturriekin elkartu gabe, emaitza hobeak lortzen baitira horrela.

Baliabideetan oinarritutako gomendioak ematen direnean, ordea, folksonomien egitura oso ezberdina da, honek distribuzioetan eragiten du eta, ondorioz, baita pisu-funtzioetan ere. Hau dela-eta, pisu-funtzioak erabiltzerakoan emaitza txarragoak lortzen dira, eta beste datu iturriekin elkartu beharra dago hobetu ahal izateko. Elkartzerakoan, baina, TFK baino emaitza hobeak lortzen ditu, sailkatzailearen iritzi egokiei esker.

### 8. Ikerketa Galdera

*Zein da etiketek kolekzioan duten adierazgarritasuna neurtzeko hurbilpenik egokiena?*

Ikertutako pisu-funtzioen artean, laster-marka frekuentzietan oinarritzen denak lortzen ditu emaitza onenak sistemak baliabideetan oinarritutako gomendioak ematen ez dituenean. Kasu hauetan, IBF da onena, IRF eta IUFk jarraituta. Horiek guztiek argi eta garbi gainditzen dute TFren errendimendua, bai bakarrik erabilia, eta baita sailkatzaile bateratuen bitartez beste datu iturri batzuekin elkartzerakoan ere.

Bestalde, guneak baliabideetan oinarritutako gomendioak ematen dituenean, erabiltzaileen frekuentziak emaitza hobeak ematen ditu. IUFren errendimendua IBF eta IRFrena baino hobe da, mota honetako guneetako gomendioetan oinarritu beharrean bere etiketa propioak definitzen dituzten erabiltzaileek duten garrantzia dela-eta. Bakarrik erabilia IUFk kasu honetan TF gainditzen ez badu ere, beste datu iturri batzuekin elkartzean emaitzarik onenak lortzen ditu. Hala ere, gutxiगतik gainditzen du TFn oinarritutako sailkatzaile bateratuen emaitza, eta bietako edozein erabil liteke emaitza antzekoak eskuratuz.

### 9. Ikerketa Galdera

*Ba al dago erabiltzaile profilak ezberdintzerik, sailkapen ataza batera ahalik eta gehien hurbiltzen diren anotazioak egiten dituzten erabiltzaileak bilatu asmoz?*

Erabiltzaile mota hori, Sailkatzaile izenekoa, existitzen dela frogatu dugu. Gure esperimentuen arabera, hau egia da, batez ere, etiketen gomendioak ez dituzten guneetan, hau da, LibraryThing-en. Gune honetan Sailkatzaileek definitutako etiketek emaitza hobeak lortzen dituzte sailkapenerako. Gomendioak ematen direnean, ordea, erabiltzaile hauek antzematea zailagoa da, GoodReads eta Delicious-ekin gertatzen den bezala. Hala ere, erabiltzaileak antzemateko neurri aproposa erabiltzeak Sailkatzaileak antzematea ahalbidetzen du, kasu hauetan ere bai.

### 10. Ikerketa Galdera

*Zeintzu dira erabiltzaile bat baliabideen sailkapen on bat egiten ari dela zehazten duten ezaugarriak?*

Aztertutako bi ezaugarrien artean, sailkapen atazarako proposenak diren Sailkatzaileak antzemateko ezaugarri interesgarriena erabiltzailearen hiztuntasuna dela erakutsi dugu. Erabiltzaile batek etiketa gutxiago edo gehiago definitzeko duen ohitura adierazten du hiztuntasunak. Ezaugarri hau baliatuz, posible da sailkapen atazatik gertuago dauden erabiltzaileak aurkitzea. Bestalde, erabiltzailearen berbategiaren aniztasunaren arabera Deskribatzaileak diren erabiltzaileak antzeman daitezke. Honez gain, etiketa deskribatzaileak erabiltzen ez dituztenek sailkapen hobe sortzen dutela deskubritu dugu.

## E.6 Ekarpen Nagusiak

Lan honen ideia berritzailea baliabideen sailkapenerako etiketa sozialak baliatzean datza. Guk dakigula, etiketa sozialak baliatuz egiazko sailkapen esperimentuak burutzen dituen lehen lana guk aurkeztutako lehena da (Zubiaga et al., 2009d). Horren aurretik, Noll and Meinel (2008a) egileek etiketa sozialak eta adituen sailkapenak alderatu zituzten analisi estatistikoa eginez. Arlo honetako lanen gabezia kontuan hartuz, tesi honetan aurkezten dugun lanak baliabideen sailkapenerako etiketa sozialen erabilpen eta errepresentazio proposerako argipenak ematen dira. Konkreteki, hurrengo ekarpen nagusiak aurkeztu ditugu lan honetan:

- Tamaina handiko 3 kolekzio sortu ditugu etiketa sozialen sistemetan oinarrituz, kontuan hartutako baliabideei adituek esleitutako sailkapen datuekin batera. 3 hauek ikerketan erabiltzeko datu-sorta handien artean daudela esan genezake eta, guk dakigula, baliabideen sailkapenerako erabiltzeko handienak dira. Datu-sorta hauetako batzuk, beste txikiago batzuekin batera, publikoki eskuragarri utzi ditugu beste ikertzaile batzuek erabili ahal izan dezaten<sup>11</sup>. Datu-sorta hauek, besteak beste, Godoy and Amandi (2010) eta Strohmaier et al. (2010b) egileek baliatu dituzte beraien ikerketa lanetarako.
- Gure lana etiketa sozialen errepresentazio ezberdinak alderatzen dituen lehena da. Gainera, etiketa sozialak eta beste datu iturri batzuk alderatuz egiazko sailkapen esperimentuak egiten dituen lehen ikerketa lana da. Etiketa sozialak goi-mailako kategoriatarako baizik, maila baxuagoko kategoria zehatzagoetan sailkatzeko ere baliagarriak direla erakutsi dugu. Noll and Meinel (2008a) egileek ondorioztatuko hipotesia ezeztatzen dugu honenbestez. Lan horretako analisi estatistikoaren arabera, kategoria zehatzagoetarako etiketen erabilgarritasuna oso txikia izan zitekeela diote egileek.

<sup>11</sup><http://nlp.uned.es/social-tagging/datasets/>

- Etiketa sozialek folksonomiatan dituzten distribuzioak aztertu ditugu, eta sistema bakoitzaren ezarpenek zentzu honetan duten eragina ikertu dugu. Horretarako, pisu-funtzio ezagun baten oinarritu gara, TF-IDF, folksonomien egitura hauetara egokituz.
- Sailkatzaile bezala definitu ditugun erabiltzaileek osatutako multzoa existitzen dela erakutsi dugu. Erabiltzaile hauen anotazioak gertuago daude adituen sailkapen taxonomikoetatik, Deskribatzaile deitu ditugun bere erabiltzaile batzuen anotazioetatik baino. Sailkatzaile eta Deskribatzaileak ezberdintzeko hurbilketak lehendik ere frogatu baziren, hauxe da Sailkatzaileak baliabideen sailkapenerako aproposagoak direla erakusten duen lehen lana.

Etiketa sozialak baliabideen sailkapenerako erabiltzea ikerketa lerro berria zen tesi honekin hasi ginenean. Hala ere, azkenaldian sare sozialetan, eta bereziki etiketa sozialen sistemetan, erabiltzaileek sortutako edukien gainean ikertzeko sortu den interesa dela-eta, lan berri ugari ekarri du. Hazkunde honekin batera, ikertzaile gehiagok erakutsi du etiketa sozialak sailkapenerako erabiltzeko interesa, eta arlo honetan egindako ikerketa lanen kopuruak nabarmen egin du gora. [Godoy and Amandi \(2010\)](#) egileek, esate baterako, etiketak erabiltzen dituzte sailkapenerako, gure aurreko lan baten oinarrituz ([Zubiaga et al., 2009d](#)).

## E.7 Etorkizunerako Ildoak

Etiketa sozialak baliabideen sailkapenerako erabiltzea oraindik ere ikerketa arlo berria da, eta lan gutxi egin da honen inguruan. Tesi honetan aurkezten den lanak ahalik eta baliabideen sailkapen zehatzena lortzera bidean etiketa sozialak errepresentatzeko modu egokia zein den argitzen du. Horrez gain, etorkizunerako ildo berriak ireki ditu.

Tesi honetan guztian zehar, etiketa bakoitza ikur ezberdin bat bezala hartu dugu kontuan, izan dezakeen esanahi semantikoa aztertu gabe. Zentzu honetan, etorkizunerako lan interesgarria litzateke analisi semantikoa egitea etiketen artean dauden sinonimoak eta erlazio ezberdinak antzemateko. Lengoia naturalen prozesamendurako teknika baliatuz, edo hurbilketa semantikoetara joz, etiketen inguruko ezagutza areagotzea lortu liteke, folksonomien azterketa sakonagoa ahalbidetuz.

6. Kapituluari erabili ditugun pisu-funtzioak testu kolekzioetan erabiltzeko pentsatutako TF-IDF funtzioan oinarritzen dira. Etorkizunerako interesgarria izan liteke beste funtzio batzuk probatzea, eta baita folksonomien egitura hauetara egokitu daitekeen beste funtzio batzuk proposatzea ere. Honek asko lagunduko luke gomendioak ematen dituzten sistemetarako, Delicious-en esate baterako,

izan ere probatu ditugun funtzioek ez baitute behar bezala funtzionatu sistema honetan.